CHAPTER I

Planning and Intelligence

The initial difficulty besetting all devisers of mental tests is that after many years of psychological discussion we have not yet formulated a satisfactory definition of intelligence. Until we can agree upon what we are attempting to measure, it is difficult indeed to discover how well we are measuring it. This does not mean that we must have clear ideas about the nature of intelligence before devising our tests. As has been pointed out, physicists can measure electricity without knowing what electricity is; ohms, amperes, volts are basically measures of what electricity does. Similarly, an operational definition of intelligence will suffice, though ultimately empirical observations should fit into a theoretical framework.

For many years I have upheld the view that an essential component has been left out of the psychological description of what intelligence does. That component is planning capacity, which seems to include prevision and prehearsal in the attainment of a goal. The failure to realize that planning is inseparable from every intelligent activity, and the inability to make adequate provision for its measurement, are two of the main reasons why the psychometric approach in clinical psychology is weak and incomplete. That approach has been hindered by considering planning and foresight as nonintellectual factors, instead of integrative parts of every intelligent activity.

In the 1920's, the heydey of the mental-testing movement, there was a great deal of discussion about what intelligence does, and how it should be measured. The most comprehensive exposition was a symposium [1*] initiated by the *Journal of Educational Psychology* in 1921. Strangely enough, it was stimulated by the wide use of group tests, which promised an easy approach to the mass

* Numbered references will be found at the end of each chapter.

determination of mental levels. That promise has not been fulfilled. Group tests are still in use, but only a hardy psychologist would dare to found any important generalizations regarding intelligence from the scores.

Some of the leading psychologists of that day contributed to the symposium and it was soon evident that intelligence was too complex a process to be covered by applying the Binet forms then available. Apparently proceeding on the principle that intelligence was what the tests measured—a definition actually proposed by one educationist—clinical psychologists began to make the approach more general by multiplying tests of all kinds and combining them into scales, regardless of the comparative validity of the individual subtests. The original Binet had been amplified into the Stanford Revision,[2] later to be further expanded into the Terman-Merrill [3] Revision, with two alternative forms. To offset the verbal imbalance of these tests, performance scales such as the Pintner-Patterson and the Arthur [4] came into use. Unfortunately, these rather hodgepodge collections paid still less attention to the comparative validity of the individual items.

Arthur, however, did weight the tests she used according to their discriminative value. This was an index of change of performance with regard to adjacent levels of chronological age, useful indeed with reference to standardization but not necessarily any better measure of validity than are changes in bodily stature. If a shotgun prescription was in order, then it should have provided as much scatter as possible. The application of the principle is recognized by the phrase "the test battery," and the Arthur scale provided a useful supplement to the verbal tests.

This unconcern with subtest validation means that, included in the psychometric armamentarium, are some weapons that are loaded with very small shot even though they add materially to the noise. But as in most mass bombardments, accurate aim is of little consequence. Credit must be given to Wechsler [5] for modifying the confusion of multiplying and assembling tests. In his verbal scale the tests are more thorough and detailed in application and thus more concentrated in effect. His performance scale covers an entirely different field of fire but, unfortunately, very inadequate proofs of subtest validity are forthcoming. In spite of the almost general acceptance of the Wechsler-Bellevue scale, I believe that the target territory is inadequately mapped and that important sectors are out of range.

Continuing the analogy, we may say that the 1921 reconnaissance brought back considerably more information than the test devisers

have used. Particularly was this the case with regard to planfulness, which either by direct statement or implication was included in almost every attempt to define intelligence by what it does. The objective was clearly marked down but there were only sporadic and ill-supported advances in its direction.

Thorndike,[6] for example, in his much quoted contribution to the symposium, divided intelligent behavior into two classes: (1) responses to situations directly sensed, and (2) responses that are primarily the result of planning. But though he clearly stated that planned responses call for greater intelligence, he did not emphasize the obvious need to develop specific tests of foresight and planfulness in measuring intelligence.

In the practice of clinical psychology since Thorndike's day, there has come about partial realization of the inadequacy of general intelligence tests, especially of the verbal type, but unfortunately there has been a dearth of enterprise and research to fill in the gaps. I acknowledged my own contribution to this unfortunate situation when I described the Maze Tests as a *supplement* to the Binet, implying that what was needed was to strengthen the initial line of attack, instead of to plan advances on new fronts. I have also countenanced the pooling of dissimilar test scores instead of considering them separately, giving to each its diagnostic weight according to the subject's background and the purpose of the mental examination.

For example, if the aim is to predict the social adjustment of a person near the feebleminded level, then much more importance attaches to a Porteus than to a Binet score. If, however, prediction of scholastic progress is involved, then much more weight should be given to a general intelligence type of test.

It is true that the psychometric front has been greatly extended by the various predictive tests, e.g. the Bender-Gestalt, the Good-enough, the Leiter Performance, the Healy Picture Completion, the Raven Matrices, and some of the Wechsler-Bellevue Performance subtests. These certainly represent a considerable improvement in scope as compared with the Goddard and other tests, such as the Mannikin, Mare and Foal, Ship test, or the Cube Imitation, and the like. In my experience these early tests do not seem to have any great diagnostic significance.

But it is my belief that all the performance and projective tests badly need the pruning process of rigorous research as regards their validity. Unfortunately, the clinical psychologist's prestige is so adversely affected by any uncertainty in relation to his tools of measurement that he often claims much more than he is ready to

prove. His field of training is now so broad that he rarely has to answer challenges from outside his own discipline. But the absence of such attacks should not confer any sense of false security.

If the members of my own profession had declined to trust mental tests further than the validating evidence warranted, there would be no need to quote, a third of a century later, Thorndike's 1921 appraisal of the value and limitations of the psychometric devices of his day. Unfortunately, his strictures are relevant today. He wrote:

> If the boy has had ordinary American opportunities, his score (in standard tests of the Binet type) will prophesy rather accurately how well he will respond to intellectual demands in case of "book-learning" at the time and for some time thereafter, and, very possibly, for all his life. It will prophesy less accurately how well he will respond in thinking about a machine that he tends, crops that he grows, merchandise that he sells, and other concrete realities that he encounters in the laboratory, field, shop and office. It may prophesy still less accurately how well he may succeed in thinking about people and their passions and in responding to these.

In Thorndike's view, therefore, thinking falls into three categories: abstract or verbal, practical or mechanical, and social. Whether goodness of response (to use his criterion of effective thinking) in these different fields means that there are different kinds of intelligence is beyond determination at present. No one can say whether the abstract intelligence, which Einstein exhibited, would have been equally effective if he had devoted the same time and mental energy to practical inventions, or had shown as much interest in literary as he did in mathematical symbols. Obviously, such questions are beyond the range of experiment.

The whole matter would be of academic interest only, except for the fact that the main value of psychometrics to clinical psychology depends on accurate prediction. Practically speaking, clinical psychology is much less concerned with measuring or assessing genius than it is with measuring or assessing persons at the other end of the intelligence scale. To grade mental tests at the average or subnormal levels should not be a too difficult task, whereas to apply psychometry to genius may require the assumption that the examiner himself is a genius. It would seem, therefore, unwise to pay too much attention to the assessment of the highest abstract ability, since only a psychological Einstein or Shakespeare could have devised tests that would have predicted the effectiveness of geniuses. In the absence of triangulation data, the simplest judgment of comparative altitude is made from similar heights.

On the other hand, the adjustments to living commonly called for, such as the growing of crops, the sale and manufacture of goods, construction of all kinds, and even the management of people, may well come within the scope of predictive psychometrics. These adjustments obviously call for the exercise of planfulness, by which I mean foresight and the tendency to use it.

Terman,[7] in his contribution, defined intelligence as the ability to carry on abstract or conceptual thinking, but this statement is frequently quoted out of context. As a matter of fact, he gave full weight to the imaginative manipulation of perceptual data in the evolution of concepts. In describing the intelligent man, he said: "Add (to sensori-motor and perceptual levels of thought) a large measure of ability to associate abstract ideas into complex systems, and he can design a new type of engine, draft the plans for a sky-scraper, or discover a curative serum." This description of ability for design, for drafting plans, for scientific discovery, and for effective association of ideas is merely bringing abstract intelligence down to earth, and Terman's whole statement provides excellent support for the belief that planning is inseparable from every intelligent activity. It is this recognition by Terman of the need for tests of "practical intelligence" that formed the basis of Spearman's remark that though the former defined intelligence as the ability for abstract thinking, this definition seemed to have little bearing on his choice of test items for his mental scale. It is fortunate for psychometrics that this was the case, since so many of the measures well loaded with *g* are dependent upon educational experience, and are thus inapplicable to people in the lower social and industrial levels. The only abstracting ability needed by the ditch digger is the ability to abstract dirt from a hole in the ground.

Other contributors either stated or implied the integration of planning capacity with intelligence. Freeman,[8] an authority in the field of mental measurements, sought concurrence with his definition by expressing it in the most general terms as "a balanced and sane reaction to the entire world of things, ideas and persons." The difficulty here is, of course, to say what a sane reaction is. In the thinking of many people, "sane" merely means "conventional." However, Freeman wisely goes on to suggest testing the most significant aspects of experience, nonsuggestibility, and the ability to adapt means to ends—all characteristics of the man who plans well.

Thurstone's[9] formulation was apparently founded on broad neurological considerations. In his view, intelligent action is characterized by three stages. The first is the inhibition of the immedi-

ate instinctive reaction; the second is the adjustment in the light of imaginally experienced trial and error; the third is the use of this final adjustment to the advantage of the individual as a social animal.

The relevancy of this description is evidently dependent on the neurological view of the cortex as the organ of delayed choice. Inhibition of the immediate instinctive response becomes appropriate through the development of the distant receptors of smell, vision, and hearing, which makes anticipatory delay not only possible but advantageous. Vision, because of its role in pin-pointing the stimulus in location and relevancy, becomes the greatest factor in cortical supremacy. It provides time for appropriate planning.

It is this "imaginally experienced trial and error" process that I characterize by the term *prehearsal.* Since the situation may present new elements, this mental essay, or search for a solution, is not in the nature of a rehearsal, a term applicable only to the repetition of a previously experienced response. The manipulation and rearrangement of familiar elements and their experimental synthesis with the new elements before embarking on overt action seem to me to be better described as prehearsal. It is a most essential part of planning.

Thurstone's third stage—adjustment of the response to the advantage of the individual as a social animal—might be summarized as the capacity for socially orientated, prudently planned reactions. It seems to put the necessary emphasis on the value or relevancy of the response. The high place that Thurstone assigns to planning and foresight is indicated by his statement that the more deliberative the characteristic reactions are, the more profoundly intelligent the person is.

It seems clear, therefore, that the leaders in psychology who contributed to this symposium all recognized the importance of practical foresight, but, unlike the men who "builded better than they knew," the early test devisers had wiser theories than practice.

In 1933 [10] I proposed as a definition of intelligence "the capacity to respond to a wide range of relevant stimuli." In a book published in collaboration with Wood Jones in 1928,[11] we defined the relevant stimulus as the one to which it is biologically advantageous for the organism to respond. As man is a social being, biological advantage covers both individual and social purposes, without putting exclusive stress on either, thus recognizing the fact that society and its temporarily accepted conventions may be wrong.

The relevant stimulus was illustrated by the splash of a stone in a pond, a signal that alerted all the frogs on the bank to be in readiness to dive. Another example was the effect of the sound of

falling rain on leaves in the case of the English blindworm. Though seemingly deaf to all other sounds, this was the signal to the blindworm that, with the shower, the little gray slugs on which it feeds will be abroad. To fail to respond to that stimulus meant extinction through starvation.

One advantage of this proposed definition is that it can cover the whole range of intelligence, both animal and human. The monkey finds a wider range of stimuli to which it can respond than does the dog; the wise man discovers many more relevant stimuli than the fool does. At the same time selectivity is implied, so that it is not the number of stimuli responded to that distinguishes greater intelligence but the relevancy of the selected stimuli to life's purposes.

Man's supremacy over the rest of Nature is apparently related to his lack of highly developed specialization except in one single physical feature. In everything but the cerebral cortex he has kept to the evolutionary middle of the road. He occupies a moderate position in the scale of size, strength, mobility, endurance, sensory perception, defensive or offensive equipment. Most of the lower animals excel him in one or the other specializations for such purposes. His distinctive specialization is the extreme development of the cerebral cortex, and nowhere does this seem more outstanding than in the frontal lobes. The failure of psychologists to identify a whole range of important tested abilities with the functions of the frontal lobes will be discussed later in this volume.

Further consideration has convinced me that in behavior an essential difference between man and his fellow creatures lies in his capacity for planning and that, therefore, this should be recognized in the definition of intelligence. The work of bees, wasps, spiders, and similar organisms, exhibits technical excellence, but it is doubtful whether any of them exercise either foresight or hindsight in execution. Since the design never varies, it is probable that the worker has no prevision of what the comb or the web will look like when completed, nor does the creature work back to improve or repair the structure if this is accidentally damaged or weakened. Apparently, there is no individually preconceived plan; the insect merely follows a reflex blueprint impressed on its nervous system through the experience of the species, and prehearsal is unnecessary.

From the highest animals to whom may be accorded some measure of foresight, it is still a far cry to some primitive human picking up a nodule of stone and recognizing that it would fit into his grasp to serve as a rude tool. Such an occasion must have marked the

moment of prehearsal, the mental representation of the utilitarian act, when hand and stone were united in the process of tool-using. From this concrete experience sprang the abstract concepts of size, weight, shape, and sharpness that made possible the choice of similar implements and, later still, the ideas of chipping, flaking, and grinding operations that would fit the tool better for service.

This was man's first and perhaps greatest technological discovery. Next came, through experiment, the generalization that certain types of rock shaped more easily or more suitably than others. Through this merging of perceptual and conceptual prevision came about more definite prehearsal, and man the artificer was in operation.

I am one of probably very few psychologists who have been fortunate enough to sit at the feet of a contemporary Stone Age worker while he fashioned a spearhead out of quartzite, his accessories being a wooden block, a wad of paperbark as a shock absorber, and a kangaroo bone as a tool.[12] By carefully applied pressure he flaked off small chips of stone to shape and sharpen the spearhead.

In watching the work it was easy to infer the workman's constant reference to a mental blueprint, and to observe his prudent preconsideration and prehearsal of every step in the process. As the weapon took shape and as the finishing touches of flaking off the tiny serrations on the edges of the sharp point were called for, there was evident the most meticulous hindsight and foresight, the point being held up and minutely inspected before removing the next tiny chip of stone. Any imprudent or ill-judged pressure at this stage, no matter how slight, might spoil the whole venture. Nothing less like unplanned instinctive working could be imagined. The observer could not fail to conclude that here was an artist in stone, a planful being.

Hence, it should have come as no surprise to find that these aborigines could develop artistic ability in other media. Twenty years later, eleven members of this tribe were painting water-color pictures with fine competency. The leader of the group, Albert Namajira, became one of the most popular painters in Australia, and his work is well known abroad.

Incidentally, the psychologist was able to watch the same Stone Age workman threading his way through a maze test, and exhibiting an equal degree of prudence and foresight in his performance. One naked Luritcha * native, who had never previously laid eyes

* Luritcha is the name given to a group of tribes, or subtribes, which inhabits country about 200 miles southwest of Alice Springs in Central Australia. This was near the area in which I worked in 1929, with the support of the Australian National Research Council.

on pencil or paper, obtained a perfect score in the Porteus Maze. This was a task that the natives accepted as a worthy challenge to their ability. In another part of Central Australia, Fry and Pulleine[13] applied the Maze test to aborigines of a different tribe and they remark: "While watching a native at work on these problems, one can almost feel the intensity of the mental strain involved."

While engaged in similar work in the Kalahari desert of South Africa in 1934,[14] I was amazed at the Bushmen's extreme interest in the same test. The habit of decorating shields, totemic objects, and even their own bodies with geometric designs, which is characteristic of the Australian aborigines but not of the African Bushmen, whom I examined, may have helped to account for the superior Maze Test performance of the former. There was, however, no difference to be noted between the two groups in degree of interest in the tests.

Since the process is unconscious, we are unaware of the extent to which anticipatory prehearsal underlies all purposeful reactions. For example, there is the experiment of measuring electric potentials set up in the muscles of the neck and shoulders when subjects are asked merely to imagine they are planning to look at the ceiling of the room in which they are sitting. The changes in potential reveal the nervous and muscular anticipatory set that takes place unconsciously, even if the plan is not carried out. I have suggested that reaction time has little to do with the speed of thought, but represents the time necessary to set up this complex anticipatory readiness to respond.

In mental tasks, such as making an impromptu speech, we are often conscious of the laborious and even distressing process of what Thurstone describes as "cutting and trying with alternatives that are not yet fully formed in behavior." How much more complicated is this choice of alternatives in complex scientific problems! Titchener once suggested that the good research worker is a man who is fertile in expedients. We are all familiar with this process of choosing, trying, and rejecting or adopting alternative courses of conduct or thought. At a simple level this is similar to solving a very complex maze. There is even an analogy between the Maze and the cortex itself, with its tangled mass of axons and dendrons as nerve pathways and its millions of synapses as choice points.

Enough has been said to indicate that planning enters into every intelligent act as an essential component. If that is so, it suggests that some specific measure of this capacity should be included in any adequate measure of intelligence. Later in this volume some

evidence will be provided that the frontal cortex, particularly in its superior areas, is related to planfulness.

Throughout this book the major emphasis will be placed on the validity of tests, particularly the Maze. The failure of test devisers to provide adequate evidence with regard to validation is due not so much to a misdirection of effort as to the inherent difficulties of the problem. There is rather general recognition that the validating devices commonly employed are imperfect. The method of group comparisons, consisting of the mean results of tests that are to be used for the mental diagnosis of individuals or for individual prediction, is a case in point. Another ill-founded procedure is the attempt to establish the validity of a measure through its agreement with other tests, themselves imperfectly validated. Most reprehensible of all is the bland assumption that intelligence is what the tests measure and that, therefore, the affixing of the label "mental test" is equivalent to setting up an acceptable claim for validity. Stating also that the test "works" is little more than saying that thousands of psychologists cannot be wrong and that their intuitive judgment that the test is a good one can be substituted for proper validation by rigorous research. Then, too, some test devisers condemn the validity of teachers' judgments but do not hesitate to appeal to this criterion.

The great deficiency that affects the proper interpretation of all test verdicts, including the Maze, is the absence of any satisfactory quantified measures of social behavior. If these were available, correlation studies would soon discover the comparative validity of mental tests for specific purposes.

Admittedly, to supply such measures is most difficult, and under ordinary conditions an almost impossible task. The research psychologist has no authority to go out into the highways and byways and gather in his subjects for examination. Hence, he is dependent upon partially controlled situations such as the schools provide. Thus, his work is likely to be confined to children in the primary and high school grades, who in the latter case are often educationally selected and above the average in intelligence. The tests evolved by such applications are likely to be too literate in content and too difficult for the ordinary population.

As regards truly unselected subjects, the difficulty is to make them stand still long enough for careful, detailed observation and examination. So it comes about that a simple physical measurement such as stature is a comparatively easy matter to investigate, but the systematic observation of social behavior is a lengthy and sustained process. Even if the work were done in the controlled

environment of a public institution, the research plan as initially set forth could rarely be adhered to exactly. Any psychologist who has done research in a mental hospital, for example, is painfully aware of the forces of attrition that often diminish the size of his experimental or control group. Having little or no authority over his subjects, his research is ordinarily dependent upon the good will of supervisors, nurses, aids, attendants, as well as doctors and superintendents, to say nothing of the cooperation of patients.

It would seem necessary, therefore, to go back to much earlier procedures when the best work on the validation of mental tests was being carried on in institutions for the feebleminded. Though the results obtained there may not have general application, they would have a direct bearing on the problem of mental diagnosis.

A methodological difficulty in setting up a social criterion based on nonintellectual attributes is that of semantics, which often compounds the fallibility of human judgments. But any such criterion must be nonintellectual in content; otherwise, we shall lapse into the circuitary process of proving the validity of the tests by the tests themselves.

Yet the obvious advantages of establishing an external criterion are tremendous. If we could only declare, with any measure of assurance, what a Binet, Wechsler-Bellevue, Porteus Maze, or Rorschach test score means in terms of social adjustment or adaptability, then psychometry would no longer depend on the shifting sands of teachers' judgments, cross-validation, examiners' intuitions, psychologists' fashions, popularity, and the like.

The provision of such a court of appeal is, to my mind, the greatest service that could be rendered, not only to psychometrics but to psychotherapy. Its lack is responsible for most of the shadows that envelop the practice of clinical psychology today.

Chapter References

[1] "Intelligence and Its Measurement: A Symposium," *J. Educ. Psychol.*, **12**, 123–47, 195–216, 271–75 (1921).

[2] Terman, Lewis M., *The Measurement of Intelligence* (Boston, Houghton Mifflin Co., 1916).

[3] Terman, Lewis M., and Merrill, Maude A., *Measuring Intelligence* (Boston, Houghton Mifflin Co., 1937).

[4] Arthur, Grace, *A Point Scale of Performance Tests* (Commonwealth Fund, N.Y., 1930, 1933).

[5] Wechsler, David, *The Measurement of Adult Intelligence* (3d ed.; Baltimore, Williams and Wilkins, 1944).

[6] Thorndike, E. L., "Symposium," *loc. cit.*

[7] Terman, Lewis M., "Symposium," *loc. cit.*

[8] Freeman, F. N., "Symposium," *loc. cit.*

[9] Thurstone, L. L., "Symposium," *loc. cit.*

[10] Porteus, S. D., *The Maze Test and Mental Differences* (Vineland, N.J., Smith Printing and Publishing House, 1933).

[11] Jones, F. Wood, and Porteus, S. D., *The Matrix of the Mind* (Honolulu, University of Hawaii Press Association, 1928).

[12] Porteus, S. D., *The Psychology of a Primitive People* (London, Arnold & Co., 1931).

[13] Fry, H. K., and Pulleine, R. H., "The Mentality of the Australian Aborigine," *Aust. J. Exper. Biol. and Med. Sci.*, 8, 153–67, 1931.

[14] Porteus, S. D., *Primitive Intelligence and Environment* (Macmillan Co., New York, 1937).

CHAPTER II

Psychometric Diagnosis

Goddard years ago pointed out the advantages of studying the retarded, likening their mental development to the passage of a slow freight train, rendering it an easy matter to decide what it carries. So, too, the difficulties involved in attempting to measure intelligence should lessen as we approach its lower limits. This chapter, in partial contrast to the previous chapter, devotes attention to the activities of the unintelligent.

Diagnosing feeblemindedness is naturally one of the main concerns of clinical psychologists, since the psychometric means of doing so have been devised by them. Also naturally, the validation of such tests has been left in their hands. Latterly, the recognition of children with very high productive ability has become of extreme psychological interest, though the problem of knowing what to do with them is the educationist's responsibility. This concern with gifted children has been heightened by the strong suspicion that the American educational system is inadequate for present-day requirements. It would be odd indeed if Russian rivalry in science forced attention to the traditional need in education—hard work for both teacher and pupil. This would also react on psychological practices, particularly on measuring nonintellectual factors.

So far, the selection of children with high intellectual and scientific potentiality has been hampered by inadequate definition of intelligence and inadequate means of predicting productivity. Generally speaking, we have been plagued by the fact that our measures have been too verbal or literate and insufficiently practical. If it is true that our educational system is faulty, then tests that are over-scholastic may mean that we shall find ourselves backing away from the real problem.

More than ever the world needs what Freeman called "a balanced and sane reaction to things, ideas and persons," and our present way of life is a poor reflection of such an attitude. To at-

tain a better balanced and saner society calls for more purposeful efforts and greater self-sacrifice than most of us are willing to expend. At the beginning of the present era of clinical psychology, Binet [1] emphasized as a guiding principle a testing aim that was more practical than scholastic. In 1905 he wrote: "It seems to us that in intelligence there is a fundamental faculty, the alteration or the lack of which is of the utmost importance for practical life. This faculty is judgment, otherwise called good sense, initiative, the faculty of adapting one's self to circumstances. To judge well, to reason well, these are the essential activities of intelligence." But if this is true, the psychologist's repertoire is not overstuffed with tests of adaptability, good judgment or practical reasoning.

Unfortunately, generations of extolling literary and artistic talent have fostered the notion that its productions should be exempt from common-sense standards. The highly intelligent man must learn to accept more eagerly the role of exponent of scientific truth, not merely the titillater of the lighter emotions of mankind, no matter how rich the rewards of such titillation. For the time being, the world apparently has enough poets, painters, writers, musicians, and entertainers, but needs desperately more real scientists. But even science may not be an end in itself, and in selecting the educational elite we must learn to recognize that possession of common sense and good social judgment should be one criterion, an unfailing accompaniment of scientific know-how.

Such a discussion of philosophic values in life and education would be extraneous here except for the fact that one function allotted to psychologists is to assist in removing the grit from the world machine. The early recognition of the socially and mentally inadequate is the necessary complement of our efforts to detect those whose high talents will make life better and safer for the rest of us. Unfortunately, clinical psychologists have not yet succeeded properly in differentiating the unfit and the misfits in our midst.

With regard to mental deficiency, though the responsibility for its diagnosis is shared with the psychiatrist and the social worker, the clinical psychologist carries the main burden. His collaborators are fully capable of judging the social history, but this is mainly a matter of hindsight rather than of prediction. In the first place, an adult under diagnostic scrutiny has had an opportunity to demonstrate his social inadequacy, but with juveniles such a demonstration has not often been possible. Nor is school failure always a sufficient proof of mental inadequacy. Only by properly using and interpreting objective tests can we diagnose feeble-

mindedness, except at its lowest levels where skill in diagnosis is unnecessary. Furthermore, the interpretation of tests is only possible by one who is familiar with their pitfalls and limitations, and this means complete familiarity with the research methods necessary for their devising and validation. This should dispose of the theory that the psychometrician is only a technician and that the medical man must carry on the interpretation of test results. Only in a small percentage of cases are neurological or other physical symptoms important in diagnosis, and even when they are evident, a proper psychological examination is necessary to determine their effects on intelligent adaptation. Unfortunately, some psychometricians themselves fail to meet the requirement of adequate knowledge of research in test validation. Hence the blame for diagnostic errors in the field of mental deficiency must be assumed by the psychologist. And errors have certainly occurred. In 1943, the Commissioner of Education [2] for the State of New York, and himself a psychologist, asked: "How many children and adults have landed in feebleminded institutions on the basis of a single I.Q. in the 60's, unsubstantiated by further testing or clinical corroboration, and without regard to the fact that many another of like rating was able to lead a free and useful life?" Just as is the case with recognizing high intelligence, progress in mental diagnosis has been hindered through lack of a proper definition of feebleminded-edness. We can now accumulate I.Q. test scores but they are no better than their validation, and this will not depend on mathemathical criteria but on clear ideas about what we are trying to measure and evidence on how well we are succeeding in measuring it.

There have been many definitions proposed with regard to mental defect, but though these terms may vary there seems to be an underlying agreement that it is a condition of social incompetence arising from arrested or retarded mental development. The insistence on the social criterion is necessary because it is the social inadequacy that brings about public institutionalization. "Only limitations in intellectual ability and social competence," wrote Kinder [3] in an article in the *Encyclopedia of Psychology*, "is common to all defectives."

Social incompetency may, of course, arise from other causes than arrested mental development. There are many people who do not manage themselves or their affairs with ordinary prudence, but who are not feebleminded. They include criminals or delinquents whose social adjustments are inferior for other reasons than

lack of intelligence. Psychotics also may require institutional care, but before their mental illness and during its intermissions they have been socially capable.

The earlier definition that I proposed [4] not only covers mental and social incompetence, but also infers an incurable condition. It also states how diagnosis is effected: "Feebleminded persons are those who by reason of permanently retarded or arrested development existing from an early age are incapable of independent self-management and self-support." A variant of this definition is as follows: "A mentally defective person is one who is unable to attain to self-management and self-support, a condition due to inferior mental development as demonstrable by tests of proven diagnostic value."

Thus the burden of proof that the tests the examiner uses are valid for diagnosis rests squarely on the psychologist's shoulders. His obligation is to explore the whole personality of the individual under examination, and he must choose measures that have a close relation to social and mental competence. This entails constant review of such studies as have a bearing on test validation. Only with this knowledge at hand can the psychologist presume to undertake, especially in the case of children, the onus of prediction.

It will be seen that the proposed definition rests on a tripod, the three legs of which are social, industrial, and psychometric. Diagnosis is by no means easy, for self-management is often a matter of degree, and self-support is difficult to assess, particularly in times of economic depression. Thus, even in things which are a matter of record there will be doubtful or borderline cases.

This is well illustrated by the Streissgath decision in Minnesota.[5] This was the case of a woman who had been committed to an institution as a mental defective. The commitment was confirmed by a district court, but an appeal was lodged against this finding in the Supreme Court of the state. Apparently the commitment had been made on the grounds of social incompetency. The woman was the mother of ten children and the family had been on relief for some years. A county social agent reported that the woman's housekeeping was poor, the meals were inadequate, the house was dirty, and the children unkempt and unattractive. But at the higher Court hearing, neighbors testified that the woman was "of sound mind" and that she was a good mother, even if a poor housekeeper. A doctor said in his evidence that in his opinion she was "a little sub-standard" but added that a lot of people were substandard without being feebleminded. He rated her at "better than a high grade moron."

The psychologist's evidence was actually indecisive. He found that the woman had a mental age of 10 years 4 months, with an intelligence quotient of 64. But this had been calculated by using 16 years as the average adult mental age, an entirely outmoded procedure. Had an average of 14 years been used as a divisor, the I.Q. would have been 73.* This case illustrates the necessity for mental examiners to keep up with the times. Some clinicians continued using the I.Q. divisor of 16 years almost 20 years after the American Army examinations of World War I had revealed that the average mental age by tests of the Binet type was 13.9 years. Using the higher divisor had, as in thousands of other cases, artificially depressed the woman's assigned mental level.

The judge's opinion of the diagnostic value of mental tests was interesting. He wrote: "While psychological tests are convenient tools for indicating mental retardation, test results should ordinarily not be considered sufficient, much less conclusive, except at lower levels, in a proceeding for the adjudication or relief from adjudication of one as feebleminded."

As a matter of fact, psychologists would agree that both examination and diagnosis on the facts advanced were certainly far from conclusive, not only because the data based on a single I.Q. were inadequate, but because the whole examining process was substandard. But multiplying tests may still be an inadequate procedure unless the proofs of their diagnostic validity are clear. Even with a more competent choice of tests, the psychometric leg of the tripod might have been a weak one, especially when the woman's intelligence level was determined by the use of verbal, or predominantly verbal, tests. Nevertheless, mental tests remain the most reliable means of diagnosis. If Judge Streissgath's decision were to stand, it would be extremely difficult to uphold any commitment of a feebleminded child except at a very low mental level.

As matters stand, commitment procedures in U.S. courts are extremely confused. Up to 1953, when Corbett and I [6] surveyed the situation, only two states seemed to rely on psychometric definition of feeblemindedness. At one time, Ohio had on its statute books two criteria that were not in agreement. The law defined a mentally deficient person as "any person found in the State of Ohio having a mental age of 12 years or less and having an intelligence quotient of 69 or less." However, even if an average age of 16 years is used

* It seems evident to me that when tests are used to designate, as in mental diagnosis, how much below average a person may be, the average test age should be the divisor employed to arrive at an adult test quotient.

as a divisor in calculating the I.Q., a mental age of 12 years yields an I.Q. of 75. Hence, a person's mental competency might depend on whether his psychometric standing were stated in terms of mental age or as an age-mental score ratio. The moral was clear. Such persons had better not be found in the State of Ohio! The latest Ohio statute sets up entirely different criteria for recognizing a mental defective, e.g. as a person "incapable of profiting substantially by any educational program" provided by a public school, or if not of school age, having an I.Q. below 50. Ohio thus substituted one area of uncertainty for another. There are many adults with I.Q.s above 50 who require, for their own good and the good of the community, commitment to an institution for the feebleminded. There are others who do not profit substantially by a public school program, but this may be the fault of the program or the teachers and not of the pupil.

As matters stood in 1953, nineteen state jurisdictions, either by specifying the use of mental tests in diagnostic examinations or by putting psychologists on certifying boards, recognized psychological criteria. Other jurisdictions, which lack statutory definitions, leave the admission standards to the institution heads, and their decisions depend largely on the results of examinations by psychologists.

In England, judging from a study by Tizard,[7] assisted by O'Connor and Crawford, there is similar uncertainty about what mental standards should govern the diagnosis and commitment of mental defectives. The Mental Deficiency Act stated that mental defectiveness means a condition of arrested or incomplete development of mind existing before the age of 18, but according to Tizard it "lays down no criteria by which mental development may be assessed." He quotes Terman's early pronouncement that "all those who test below 70 I.Q. by the Stanford Revision of the Binet-Simon scale should be considered feeble-minded," and remarks that in psychiatric writings 70 I.Q. is still widely accepted as marking the upper limit of defective intelligence. Blindly accepting the verdicts of psychiatrists who have had no hand in devising and validating the mental tests is unwise, and certainly runs counter to the view expressed here, namely, that the business of diagnosing feeblemindedness should be left to the psychologists whose responsibility is to develop the criteria. It is true that psychiatrists are merely adopting the mistakes of psychologists, who should supply the corrections. Tizard further notes that for over a period of forty years psychiatrists have inclined to the opinion that intelligence tests must be relied upon for diagnosis except in the very severe cases where testing is unnecessary; but he adds that temperament and char-

acter are considered equally important. The assessment of temperament, however, is also a psychological responsibility. As for character, which I suppose is estimated on the social record of the individual, I should hesitate to give this weight unless it can be shown to be closely related to defects of intelligence and temperament. Otherwise, we should fall into worse confusion in differentiating the mental defective from the delinquent or criminal. In diagnosis, the dependency of the social inadequacy upon the demonstrable mental or temperamental defects must be kept in view.

Tizard and his associates used the Kohs, the Raven Matrices, the Binet Vocabulary, the Porteus Maze and the Cattell Non-verbal Test (Form I.B.); their subjects were 104 adolescent and adult high-grade male defectives in an institution in England. The I.Q.s by the various tests ranged from 65 for the Cattell to 86 by the Maze. No less than 70 cases had Maze I.Q.s above 75, and 83 cases had I.Q.s over 75 by one or more tests. Tizard comments on the fact that both the test results and the findings of psychiatrists resident in the institution showed that many patients were more intelligent than the certifying orders indicated. He further suggests that "between a quarter and a half of our subjects regarded as feeble-minded could more properly be classed as dull or subnormal rather than cognitively feeble-minded." This opinion agrees with the view expressed in this book to the effect that setting up a psychometric standard, such as Terman proposed, has been responsible for unnecessary commitment of many individuals and for the characterization of thousands of others as "morons." These last individuals were fortunate enough to miss the attention of teachers, social workers, and others interested in their institutionalization. Had they been committed, their numbers would have been great enough to have enforced correction of this faulty diagnosis.

Tizard also points out that setting up a statistically based definition of mental deficiency, as for example those individuals whose I.Q.s are below a level 2½ times the S.D., must take into account that standard deviations differ widely in the various tests. For example, the 1937 Stanford Revision of the Binet has larger S.D.s than the 1917 form "so that about ten times as many adults would have I.Q.s as low as 70 on this as had I.Q.s of 70 on older versions of the test." [8]

Unfortunately, there is no marked agreement among psychologists about what mental deficiency is nor how it should be measured. Wechsler,[9] for example, has proposed that persons be considered mentally defective if their Wechsler-Bellevue scores deviate –3 PE

from the mean. This, he states, would mark off 2.15 per cent of the general population as feebleminded. In 1952, Weinhofen,[10] on the basis of a survey, asserted that there were 110,000 mental defectives housed in institutions for the feebleminded throughout U.S.A. If Wechsler is correct in his general estimate, out of a population of 160 millions, there would be about 3½ million mental defectives in the country. That is to say, for every feebleminded person committed to an institution, there are 31 at large in the community. In my opinion, either the percentage as estimated by the Wechsler-Bellevue cut-off is far too high, or else feeblemindedness is far from being the social detriment it is thought to be. My own estimate of the incidence of mental defect would be a much more moderate figure—about 0.5 per cent of the population. Even that conservative figure would total 800,000 feebleminded, which would still mean that for every defective institutionalized there are seven at large.

Wechsler also states that "there are individuals with I.Q.s below 60 whose entire life history is that of a non-defective," and that there are others with I.Q.s above 70 whose behavior "is definitely defective." This statement in itself illustrates clearly the futility of a percentage definition.

Wechsler cites the case of an individual with a Stanford-Binet of 13 years 8 months (T.Q. 91) who could not hold a job and who had been arrested for sex misbehavior. The implication was that he was a mental defective. There are many idle and dissolute persons who could be classified as morally and socially inadequate, but who are not, by the definition accepted in this book, feebleminded. They are socially incompetent, being neither self-managing nor self-supporting, but their condition does not arise from mental arrest or permanent retardation.

With other pronouncements by Wechsler, I am in complete agreement, particularly as regards his protest against the idea that performance tests are relatively less diagnostic than verbal measures are. "Our scale," he said in the 1944 edition of his book, "assumes the contrary. It was constructed on the hypothesis that an individual manifests intelligence by his ability to do things, as well as by the way he can talk about them." [11]

Many years earlier I expressed the same idea when I wrote: "After all, in everyday life, what a person knows is not a sufficient indication of what he is. What he can do provides a much more practical test of his mental worth." [12]

Wechsler's definition of intelligence is apt and comprehensive: "Intelligence is the aggregate or global capacity of the individual

to act purposefully, to think rationally, and to deal effectively with his environment." This is an excellent operational definition, but it is questionable whether his scale is a sufficient measure of purposeful, rational, and effective action, except perhaps as regards its verbal components. In the shotgun union that he has brought about between verbal and performance tests, it may be that some of the last named have married above their station. It is easier to devise tests that are really indicative of what a person can say than it is to find adequate measures of what he can do. Practical ability has so many more facets than language ability that it is more difficult to select performance tests that are diagnostic.

For example, reassembling the dissected parts of a pictured mannikin, a head profile and a hand would seem to have a rather tenuous relationship to practical ability and an even more remote relation to artistic or creative sense. The author of the Wechsler scale confesses to much hesitation about including the Object Assembly test in his scale. The Block Design test also has some excellent features, especially its relation to design, but its main fault is that it is a purely imitative task. Wechsler quotes Nadel's findings that success in this test is affected by frontal lobe lesions. If this observation is valid, it is very strange that the Kohs test had, according to Hunt's study,[13] a very inconsistent relation to the most drastic frontal lobe lesion, lobotomy. She reports that 33 per cent scored better after operation, 30 per cent remained the same, and in only 37 per cent the scores declined.

Similarly, seeing the missing features in drawings of familiar objects, such as the absence of one leg from the drawing of a crab, seems to have little relation to everyday adjustments and practical ability. That kind of perceptual analysis is too limited and specific. The gravest criticism of Wechsler's Performance scale is that it fails to examine planfulness, the essential lack of the scale as a whole. Thus, one criterion that Wechsler uses in determining a subtest's worth, namely, high correlation with the rest of the tests in the battery, is a detriment rather than an advantage. If the field of fire is too limited, then obviously to include a test because it has a similar short range is faulty methodology. As I have contended elsewhere, all intelligence tests to some extent involve planning, or prehearsal, but they test so many other intellectual abilities that planning is lost in the shuffle. It would seem most desirable to test this capacity in specifically recognizable operation.

With regard to the Wechsler-Bellevue Test, Tizard questions its standardization for an English population and also the "face validity" of some of the items of his Performance Scale. I am in full

agreement with this position as well as with the statement that "no test at present in use can be considered satisfactory" for diagnostic purposes. In my opinion, no diagnostic examination should be considered complete without some measure of planfulness as an essential factor in intelligence.

It may be well to point out that I am by no means alone in my general criticisms of scales such as the Wechsler and Binet; the chief difference is that my own objections are of considerably longer standing. Cronbach,[14] for instance, lists some of the defects of the Binet as follows:

1. The scores are affected by previous educational experience and therefore do not reflect native capacity.
2. The tests are too verbal.
3. Culture differences affect scores.
4. Scores at different levels represent different mental abilities.
5. The varying sizes of standard deviations affect the distribution of I.Q.s.
6. Emotional attitudes influence results.
7. Specific mental abilities cannot be identified.

The last named objection is stressed in this volume, although planning capacity is more general in its scope than specific.

Wechsler's criticism of the Binet is no less trenchant, while Stoddard goes so far as to say that as a diagnostic instrument, the Binet is neither very reliable nor very close to a measure of intelligence.

Though it will be seen later that through the unexpected provision of a striking and new criterion the Maze has received almost unique validation, this by no means infers that it can stand alone. On the basis of the belief that normal persons require all-round ability not far below the average twelve-year score, there is still an urgent need for the Binet and Wechsler type test scales, which involve a wide range of mental abilities, and the wider the scatter the better. But a necessary component of such a basic intelligence index is a measure of planfulness. My attitude toward the value of tests is made clearer by an analogy that I recently employed:

> The task assumed by clinical psychologists and psychiatrists in the assessment of human personality is analogous to an attempt to discover what is going on inside a large building to which we have no key. We are, however, permitted to peep in any windows that are within reach. Obviously, to look through a single window would give us a most inadequate viewpoint.
>
> To my notion, tests such as the Binet give us access to a whole row of windows along one side. The Wechsler-Bellevue provides similar

opportunities with the addition of a few rather unclear performance tests that serve as windows along another side. The difficulty is that most of the general intelligence windows seem to open into a library that is not in frequent use. Projective tests also provide another set of apertures but the panes badly need the cleansing touch of much validating research. On another side of the building which should give visual access to rooms where much of the planning of work goes on, there seems to be only one small opening, the Porteus Maze. It is high time that other and better tests of anticipatory reactions and planfulness are provided, as well as a thorough window-cleaning job on those we already have.

It is further suggested that the Maze window looks in on the main living level.

Chapter References

[1] Binet, Alfred, and Simon, Th., *The Development of Intelligence in Children* (Trans. by Elizabeth S. Kite) (Vineland, N.J., Research Dept. Publications, 1916).

[2] Stoddard, George D., *The Meaning of Intelligence* (New York, Macmillan Co., 1943).

[3] Kinder, Elaine F., "Feeble-Mindedness," *The Encyclopedia of Psychology* (Ed. by Philip L. Harriman) (New York, Philosophical Library, 1946).

[4] Porteus, S. D., and Corbett, Gerald R., "Statutory Definitions of Feebleminded in U.S.A.," *J. Psychol.*, **35**, 81–105, 1953.

[5] *Re* Masters, 216 Minn. 553 (1944), 13 N.W. 2nd 487, reported in 158 *Am. Law. Reports*, p. 1210.

[6] Porteus and Corbett, *loc. cit.*

[7] Tizard, J., O'Connor, N., and Crawford, J. M., "The abilities of adolescent and adult male defectives," *J. Ment. Sci.*, Oct., 1–19, 1950.

[8] Tizard, *loc. cit.*, p. 14.

[9] Wechsler, *loc. cit.*, p. 40.

[10] Weinhofen, H., "Hospitalizing the Mentally Ill," *Michigan Law Review*, **50**, 837–72 (1952).

[11] Wechsler, *loc. cit.*, p. 137.

[12] Porteus, S. D., "Mental tests for feebleminded: A new series," *J. Psychoasthenics*, **19**, 200–213, 1915.

[13] Hunt, Thelma, in Freeman, Walter, and Watts, James W., *Psychosurgery* (Springfield, Ill., Chas. C. Thomas, 1942), Chap. XIV.

[14] Cronbach, Lee, *Essentials of Psychological Testing* (New York, Harper, 1949).

CHAPTER III

The Maze Test—Early History

Since fuller accounts embodying the more important studies are already available, only the briefest historical summary of early work with the Maze Test is necessary here, and it is included partly to show the psychological climate that affected its development and use. Like the Binet, its inception came about through the need for a better classification of mental defectives in relation to their training or education.

Before 1913, the work of Itard, Seguin, Montessori, and Witmer had brought about a more optimistic attitude toward the educational improvement and, possibly, the cure of extremely retarded children. Special schools and classes were being set up for their treatment, clinics and research departments for their diagnosis and classification. Binet, Goddard, Decroly, Degand, and De Sanctis were the leaders in the latter endeavor, to be followed later by Healy, Doll, Terman, and others. The World War I army examinations gave the work of intelligence measurement a new direction, and by 1920 the mental test movement was in full swing.

In 1913, while the chief emphasis was still educational, I was given the responsibility of organizing the school training of defectives in Melbourne, Australia, including the mental testing and selection of feebleminded children. In the very early stages, the Chief Medical Officer, Dr. Harvey Sutton, was my main collaborator, but soon his other duties removed him from active participation in the program.

The Goddard Revision of the Binet, with instructions by Edward Huey, was the best available means of mental examination, and on the basis of these tests about 140 children were selected for special education. It was soon apparent that determining the Binet age, though extremely useful, was not adequate for predicting even scholastic progress. Moreover, it had little value as an indication

of practical trainability. The standing of pupils in classes in simple toy-making, wire work, shoe-mending, cooking, housework seemed to have very little relation to mental age.

Even at that time it was apparent that planning capacity was neglected in mental measurements. Feebleminded of imbecile level were inferior subjects for education, but excellent for experiments in learning. My appointment as Lecturer in Experimental Education at Melbourne University helped to widen the field of investigation into which the experimental studies of mental defectives fitted excellently.

Work with a physically active Mongolian imbecile was interesting because the steps we took to increase the complexity of planning problems foreshadowed the methods later used by Jacobsen [1] in testing chimpanzees. (These resembled the stick and platform test, in which the animals were trained to obtain food by using short sticks to bring within reach other sticks, graduated in length, until one was finally obtained long enough to secure the food. This planning ability was later tested after the animals had been lobotomized.) Since a closed door was a constant challenge to the boy in our experiment, the chance of escape provided motivation and he was first taught how to open the door with a key. From that point progressive stages in planning were arranged, the first being to hang the key within reach on a nail. Then the key was hung in plain sight on a higher nail so that climbing on a chair was necessary to obtain it. The final training problem involved locking the door key inside a cupboard, access to which was by means of a smaller key hung again out of reach unless the subject stood on a chair. Thus the solution involved, first, climbing on a chair to get the cupboard key, second, opening the cupboard and securing the door key and, finally, opening the door. The break in the sequence, owing to the fact that the small cupboard key would not unlock the door of the room, made the problem insoluble for some time, and only the strength of the escape motivation plus frequent demonstrations made the boy persevere until ultimately he reached a solution. His persistence was remarkable, and frustration was reacted to by anger when the small key would not fit the door. Occasionally he climbed on the chair, secured the small key, then dragged the observer by the hand to the door. Had lobotomy and an adventurous neurosurgeon been available at that time so that the effect of the operation on planning could have been noted, history might have been anticipated by some twenty years. More intelligent subjects were able to get past the sequential break much

quicker, but the learning process consumed so much time that the idea of using this experimental technique to test planning in the learning process was finally given up.

The fact that some children were able to find their way to school without being convoyed by their elders offered another testing scheme. It seemed promising when it was found that one very dependable lad could read a simple sketch map of areas near the school well enough to run errands, such as delivering mended shoes to a shop in a nearby street. The idea of using a paper and pencil maze as a test of practical foresight was thus suggested. The resemblance of the test for Year IX in the present series to a section of a suburban map harks back to these early experiments.

The ability to find his way depended not only on the subject's observation, resulting in the recognition of certain cues, but also on his caution in avoiding traffic hazards. Absence of traffic lights was in those days compensated for by the slower rates of horse-drawn vehicles. Nevertheless, walking to and from a city school called for not only temperamental prudence but mental alertness or vigilance. Though a pencil and paper maze is a more intellectual task than finding one's way in a locality already familiar by experience, it does call for similar temperamental characteristics. Hence it was not surprising to find that children entrusted with the responsibility of coming to school unattended did comparatively well in the series of graduated mazes. Similarly, progress in practical work, such as shoe mending or simple carpentry, was also predictable by the Maze results since these occupations called for prudential planning. Instructors in manual work soon found that very few pupils could be entrusted with cutting and fitting together expensive material. Almost invariably, those who were able and responsible in this respect had good Maze scores.

Because it was desirable to introduce a more specific learning factor into the test situation, the method of allowing two trials on each test design was adopted and was modified to give three and finally four trials on the two highest tests.

Teaching children to write by training them to recognize basic elements in the forms of letters by tracing their outlines drew attention to the fact that the children of very low intelligence could not readily anticipate changes in direction, and this gave one the idea of drawing between guide lines as a means of extending the tests downward to the three, four, and five year levels. These designs were not mazes, since there was no choice in direction, but they did involve anticipatory reactions such as drawing between double lines forming a Maltese cross. Experiments with normal kinder-

garten children and with low grade mental hospital cases soon showed that success in these tests was valid in predicting occupational trainability.

In 1914 the most severe trial of the new test's value in mental diagnosis was undertaken with the cooperation of the teachers and principal of the School for the Deaf in Melbourne. Among the resident pupils there were about 12 per cent who in their inability to master sign language or oral communication appeared mentally backward. In addition, they made little or no progress in practical training in the School and were, therefore, considered to be feeble-minded as well as deaf. Without any prior knowledge of these children, I applied the Maze Test and was able to identify all cases classified by instructors as mental defectives on the grounds of extremely inferior learning and general adjustment.

Following this study and the presentation of its results to the 1914 meeting of the British Association for the Advancement of Science held in Melbourne,* I published articles describing the tests in June 1915 in the American *Journal of Psycho-Asthenics* [2] and the English *Journal of Experimental Pedagogy.*[3] In these contributions I stated the need to supplement general intelligence measures such as the Binet by other tests as follows: "There are certain aspects of intelligence that should be examined more thoroughly than the Binet attempts to do. These are capacities in which the feebleminded are markedly deficient. Among these we might mention mental alertness, prudence, foresight, power of sustained attention."

After stating the urgent need for tests of these capacities, I added: "Such a series should be graded for the various ages and would thus form a valuable supplement to and partial corrective of the Binet-Simon scale."

In 1915 these seemed like rash and even heretical opinions in the minds of many who were satisfied that the Binet needed neither supplement nor correction. Undoubtedly, they would have sounded better if they were promulgated by a critic who did not have a test of his own to put forward as supplying the diagnostic deficiencies of the Binet. Such statements seemed like special pleading and for some years they were regarded in the light of exaggerated claims on behalf of the Maze Test. As late as 1936, Louttit and Stackman [4] in an article on the test felt it incumbent on them to answer this criticism by pointing out that I had never suggested that the Maze could serve as a substitute for the Binet.

* The paper was presented under the title "A New Series of Mental Tests," but in the Association's *Proceedings* it was attributed to some other member.

Nowadays, the critic of general intelligence scales is no longer suspected of being unorthodox or radical in his views. At least no one leveled such accusations at Stoddard,[5] who affirmed in 1943: "Great powers and areas outside the scope of Binet testing are implied in such terms as insight, foresight and originality. . . . If, under any circumstances, mental tests fail to delineate these landmarks of intelligence, the business of identifying an obtained I.Q. with a rating of mental deficiency becomes highly impractical." Unfortunately, even with improved scales, such as the Wechsler-Bellevue, at hand, it is doubtful whether anything like diagnostic adequacy has been achieved. The Maze attempts to plug only one of several very obvious gaps.

At the time the Maze results were first published, the average Binet score of 120 subjects was 6.2 years and by the Maze 7.2 years, a difference of about 8 I.Q. points. In later investigations clinic cases in Hawaii scored about 9 points higher in the Maze than the Binet. Thus, if the two test scores were combined and averaged, the number of feebleminded on the percentage basis would be decidedly less than 2 per cent of the population, an estimate commonly accepted by psychologists. As previously stated, dissimilar tests should have their scores considered separately, though correlations of pooled test ages with a social criterion show that combined scores are more truly diagnostic.

Wechsler [6] defends the practice of considering high intercorrelation between tests as proper validation by remarking that no mental test is unique and that, therefore, a good test cannot be "markedly out-of-line" with established measures of intelligence. The truth of this statement depends on how these measures have become "established." If it is on the basis of careful validating studies that rest on external criteria, it is all to the good, but is not acceptable if the measures have been established partly on teachers' judgments and partly on popularity in clinical use. The argument that thousands of psychologists can't be wrong is of doubtful merit.

It may here be remarked that the Maze Test, in point of sensitivity to frontal lobe brain damage, comes very near to being unique. If this is so, then its disagreement with other tests significantly less sensitive might be greatly in its favor. The evidence on this point will be presented and discussed later. Wechsler himself decries validation of tests or test scales by teachers' judgments or by high correlation with the Binet, but the proofs of validity that he adduces, such as agreement with psychiatric judgments, are not any better. Actually Wechsler's citation of psychiatric opinions does not accord with his own statement that "the average psychia-

trist is no more expert in the field of intelligence diagnosis than the average teacher."

If it were true that the actual basis of any satisfactory definition of feeblemindedness is social incompetency dependent upon inferior intelligence, then it would be obligatory for any deviser of a psychometric instrument used in mental diagnosis to show that its scores are related to that criterion. I devised the Social Ratings Scale [7] in 1920 as an attempt to set up an admittedly approximate social measure that would serve as an external criterion independent of mental measurements. In its case, its weaknesses may be forgiven on the grounds that there were available neither at that time, nor since, any other means of expressing the degree of mental deficiency in terms of social behavior. Obviously, it could only be set up in an institution for mental defectives, where the judgments of people thoroughly familiar with their pupils could be relied upon, and where the experimental group was under controlled observation.

The Training School at Vineland, New Jersey, offered an excellent opportunity to work out and apply a social rating scale because of two outstanding features. The first was its active research department made famous throughout the world by Goddard's earlier work; the second was even more important, i.e., the ready cooperation of raters who had an intimate knowledge of the defectives' capabilities and social reactions. Chief among these collaborators was Alice Morrison Nash, who was responsible not only for the school training, both academic and manual, but also for arranging the weekly industrial schedules of each inmate. Her ratings were combined with those supplied by the superintendent of the home or cottage life of the experimental group, and were also based on weekly reports by industrial supervisors. Low grade cases were excluded from consideration as were also all those who had been less than three years in the institution. The following steps were taken to develop the scale:

1. Fifty terms describing the behavior traits of feebleminded were chosen from the literature on mental defect.
2. The judges took a list of 125 defectives and marked the characteristics each exhibited in terms of (1).
3. The appearance frequency of the traits was thus determined, and the eleven most commonly exhibited were chosen as trait categories for rating.
4. The 125 defectives were rated in five groups according to estimates of their comparative ability to get along in the community if they had to be dismissed from the institution. This

was called the "general social estimate." To test the reliability of these estimates the ratings were repeated within a month, using the same individuals. The correlation between the two judges' estimates was .85 and Mrs. Nash's self-correlation with a month between ratings was .87.

5. Ratings in the individual trait-categories were correlated against the general social estimates for each individual defective. Traits such as disobedience, ill-temper, and cunning or slyness were found to have negative or insignificant correlations with estimated social sufficiency, and were dropped from the scale. "Lacking planning capacity" and "lacking initiative" had such high intercorrelations that it was deemed necessary to combine them into one trait. The traits of the Social Ratings Scale now numbered seven: (1) Lacking initiative and planning (or planfulness); (2) irresolution; (3) nervous excitability; (4) impulsiveness; (5) suggestibility; (6) silliness or obtrusiveness; and (7) instability of moods.
6. These traits had finally, to be weighted and this was done roughly on the basis of the size of the individual trait correlation with the general social estimates. Ultimately, the weighting was adopted that gave the highest coefficient for the whole scale with the estimates of social fitness. This was .89 between the weighted scale scores and the individual ratings in general social sufficiency.

In order to use the Social Ratings Scale as a criterion of the validity of tests, two other sample groups were selected for rating. These consisted of 38 males and 42 females, all over fifteen and under thirty, with all really low grade cases excluded since under this condition the social ratings would be less meaningful from the standpoint of intelligence. Table 1 lists the correlations found.

TABLE 1

TEST SCORE I.Q.S AND SOCIAL EFFICIENCY CORRELATIONS

38 Males		42 Females	
Binet with Social Ratings	.60	Binet with Social Ratings	.69
Maze with Social Ratings	.60	Maze with Social Ratings	.75
Binet-Maze Aver. with Soc. Rat.	.73	Binet-Maze Aver. with Soc. Rat.	.79
Gen. Soc. Est.* with Soc. Rat.	.88	Gen. Soc. Est.* with Soc. Rat.	.89
Binet with Gen. Soc. Est.	.58	Binet with Gen. Soc. Est.	.81
Maze with Gen. Soc. Est.	.61	Maze with Gen. Soc. Est.	.81
Goddard Frm. Bd. with Soc. Rat.	.59	Goddard Frm. Bd. with Soc. Rat.	.75

* General Social Estimates

These results seemed to indicate that for these groups the Maze was at least as good a measure of social sufficiency as the Binet. Combining the mental ages or I.Q.s, however, provided a better measure than either test considered singly. The table also reveals the generally higher coefficients found with females as compared with males. This may mean that the latter are harder to rate.

The individual trait categories that correlated highest with the Maze for boys were: Lack of Planning and Initiative (.66); Nervous Excitability (.62); Irresolution (.59); and Silly and Obtrusive (.58). For girls the highest coefficients were with Lack of Planning and Initiative (.75); Irresolution (.69); Suggestibility (.62); and Impulsiveness (.58). These figures lend support to the theory that for girls the temperamental defects have a greater relation to success in the tests than they do for boys.

Dr. Marjorie Babcock, then assistant psychologist at Vineland, took the responsibility of further investigation by correlating the social ratings of other groups of defectives with both Binet and Maze tests. One very interesting phenomenon was the wide disparity obtained in the correlations according to the kind of group rated. Wherever industrial efficiency seemed the important factor the Maze advantage over the Binet seemed to increase, especially as the subjects approached normal levels. In one group only, which consisted of individuals over thirty years of age, who were rated as being of rather low mentality and had been for a very long time resident in the institution, was the situation reversed. There may easily have been a tendency to rate these old stand-bys of institution life somewhat too low, which would accord with their inferior Binet ratings. Many of them had been allotted to routine tasks, which they had fulfilled for years without change. Their static condition may have been related to institutional convenience that allowed such individuals few chances to develop initiative. The disadvantage of the Maze as compared with the Binet-Maze comparisons in the case of this group was striking enough to suggest this explanation. Table 2 gives Dr. Babcock's findings. It may be pointed out that Group VI probably represented more of a random sampling of feebleminded males since it included a wide spread of mental and industrial capacity. It will be noted that once again the Binet-Maze average gave a considerably higher correlation with the social ratings. The average Maze coefficient in all six studies was .68 as compared with .53 for Binet. I reported these results in 1922.[8]

TABLE 2

RELATION OF MENTAL TESTS TO SOCIAL RATINGS

	Group	N	Test	*r*
I	Males, good industrial workers, 14 to 30 years. Rater A.	30	Maze	.70
		30	Binet	.42
II	Same Group. Rater B.	30	Maze	.56
		30	Binet	.21
III	Males, higher mental level, 15 to 20 years, some school pupils, others industrial workers.	32	Maze	.64
		32	Binet	.62
IV	Males, high grade, 20 to 30 years, very good industrial workers.	33	Maze	.72
		33	Binet	.46
V	Males over 30 years, medium grade mentality, long resident industrial workers.	30	Maze	.78
		30	Binet	.85
VI	Males, over 14 years, wide range of mental ages.	100	Maze	.67
		100	Binet	.60
		Maze-Binet Average		.73

The Social Ratings Scale was also used in another study, which was reported in 1928. The figures concerned a wider range of tests, the subjects being 44 feebleminded males between 14 and 35 years of age, resident at the Training School for two years or longer. Cases with delinquent or psychopathic tendencies were excluded so that the group represented the well-adjusted type of institution defectives. Table 3 gives the tests used and the correlations found.

TABLE 3

CORRELATION OF VARIOUS TESTS WITH SOCIAL RATINGS

Tests Applied	*r*
1. Knox Cube-Imitation	.12
2. Goddard Form Board	.53
3. Stanford-Binet Scale	.64
4. Pintner-Paterson Scale	.66
5. Kohs Block Design	.71
6. Porteus Form and Assembling	.72
7. Porteus Maze Test	.77
8. Binet-Maze Average	.82

In respect to all the findings reported previously, it should be remembered that Vineland had no parole system, so that parents' wishes often determined the pupils' length of stay at the Training

School. There was very little follow-up of former inmates; hence, the main defect of the Social Ratings Scale was that it represented estimates and opinions that could not be checked against experience with cases paroled from the institution.

The State School at Rome, New York, was then under Dr. Bernstein who had instituted the most progressive and thoroughgoing system of paroling defectives to be found in any school for defectives in the United States. I therefore decided to check the Social Ratings Scale against actual success on parole, and accordingly transferred the investigation to Rome.

Three groups of girls were taken as the basis of this study. Group I were rated by the director of parole for their actual success in community placement, and were also scored on the weighted individual traits of the Social Ratings Scale. The correlation was .76, whereas Binet scores correlated only .32 with parole success. Unfortunately Maze scores were not available. The girls of this group had been released to the supervision of individual employers.

Group II were girls living together in an industrial colony away from the institution. The individual ratings on the weighted scale yielded a correlation of .78 with colony success, the coefficient for the Binet being .45.

Group III also consisted of 30 girls distributed between two colonies, the matrons of which relieved each other on periods of duty. Their pooled social ratings correlated .85 with colony adjustment. All the above results seem to confirm the validity of the Social Ratings Scale.

General consideration of the Social Ratings Scale reveals the fact that most of the traits, such as impulsiveness, suggestibility, irresolution, etc., are temperamental. Indeed, the only trait complex at all closely related to intelligence would seem to be planning and initiative, but from the correlations it is also evident that in the minds of the raters it is the most important. In other words, though abnormal phases of behavior, such as excitability, foolish obtrusiveness, and instability of moods are much in evidence in a feeble-minded institution, the most distinctive characteristic of the mental defective is practical inefficiency, partly through a lack of planning capacity and partly through a deficiency of the mental energy necessary to persist in efforts to solve a problem, which I have called lack of initiative. The moron's ineptitude is not revealed in routine tasks, which he may learn to perform expertly; but if the pattern or direction of the problem is changed, he fails. Thus it would seem that adaptive behavior at the level of social sufficiency depends largely on the ability to plan and execute. It will, of course,

be noted that impulsiveness, suggestibility, and irresolution are also important factors in Maze performance.

In view of the fact that my definition of feeblemindedness included under social sufficiency the ability for self-support as well as for self-management, my failure to include industrial with social ratings constituted an important defect in a scale intended to reflect the degree of mental deficiency. This lapse was needless since, with the help of Mrs. Nash, I had already arranged a work scale in which points were allotted for degrees of competency in various occupations.

The place on a ten-point scale for each job in the institution was dependent on its difficulty for a defective, especially with regard to the tools used, the value of the material worked with, the risk to the worker, and the health or safety of others. Food handling is related to the last named factors. Waiting on table is another example of high job rating, since it involves celerity and dispatch and a fairly retentive memory. The various occupational scales were themselves weighted according to the degree of individual responsibility and, conversely, the amount of supervision necessary. Thus fruit growing was heavily weighted because its operations call for independent judgment, since the supervisor cannot direct immediately each operation, and mistakes in the worker's judgment would be expensive or dangerous. Hence its comparative weight was 12 as against 7 for housework or basketry, where the supervisor exercises much more direct and constant direction of the task. As a sample, the scale for Orchard and Vineyard work is given in Table 4, showing the steps and the comparative industrial weightings.

TABLE 4

INDUSTRIAL RATING SCALE—ORCHARD AND VINEYARD

Operation	Rating	Comparative Weighting
Gathering fallen fruit	1	10
Hoeing around trees	2	20
Picking gooseberries	3	30
Picking grapes or raspberries	4	40
Sorting picked and thinning unripe fruit	5	50
Picking apples, spraying small fruit	6	60
Picking peaches	7	70
Pruning	8	80
Spraying vineyard	9	90
Spraying orchard	10	100

Obviously, picking apples can be closely supervised, there is little risk, and errors in judgment are soon apparent. Operating a horse-drawn spraying machine calls for good judgment, thorough work, and any inefficiency could affect the future crop. The full industrial rating scale is to be found in the 1922 publication previously cited. The above descriptive details are included here so that if the writer's suggestion that feebleminded institutions still present the best opportunities for validating tests is accepted, the place of industrial competency in individual evaluations will not be overlooked.

Interesting correlations with mental tests were obtained that provide additional evidence that the Maze does reflect industrial competency, at least at an average level. Vineland experience indicated that steps 9 and 10 in the scale could not be reached except by individuals at normal or near-normal levels of competency. Hence correlations should be meaningful for such single scales as woodwork for boys and needlework for girls (in which latter job, making a simple garment without immediate supervision was the highest step). Besides these correlations, others were worked out to combine the weighted scores in a variety of occupations that yielded an "Industrial Index." The coefficients are given in Table 5 for the single scales of both woodwork and needlework and for

TABLE 5

CORRELATION OF MENTAL TESTS WITH INDUSTRIAL EFFICIENCY INDICES

Test	Woodwork 43 boys	Needlework 43 girls	Industrial Index 41 boys	Industrial Index 35 girls
Binet	.60	.84	.62	.69
Porteus	.72	.88	.80	.81
B.-P. Aver.	.78	.92	.81	.83

general industrial indices. It is plain from these figures that the Maze has the closer relation to industrial efficiency *at these levels* as compared with the Binet, the girls having slightly higher correlations than boys. It is equally evident that a better index of degree of mental deficiency would result from combining industrial competency and social adjustment data, since both together reflect ability for self-management and self-support.

A noteworthy attempt to set up a measure of social independence for normal subjects was Doll's Social Maturity Scale. At younger age levels its data would be very useful for test validation purposes.

It seems to me to present two difficulties. The first is that scores depend on parents' statements and not, as in the Industrial Ratings Scale, on actually demonstrated competency; the second is that in the upper years it tends to measure the opportunity to display maturity. Nowadays, with so much more freedom allowed teen-agers, the steps of the Scale might need revision.

For historical interest we may cite two cross-validation studies reported about the beginning of the 1920's. The first was a study by Poull and Montgomery [9] of children sent for a variety of reasons to Randall's Island, New York. Its avowed purpose was to examine the claim that the Porteus Maze was "a measure of planning ability in a restricted situation," and that its scores are lower in cases of delinquency. The authors compared Maze and Stanford-Binet scores of two well-defined groups: conduct disorder cases and socially adjusted cases. There were 72 in the first group who were so classified because of "stealing, obstreperous behavior, truancy, fighting, and inability to get along with companions or guardians." The control group of 81 cases consisted of "law-abiding, self-controlled, diligent, and cooperative individuals."

The Binet comparisons showed that the socially well-adjusted group scored *lower* than the maladjusted cases, the mean for the former group being 8 years 5.09 months, while the delinquent group averaged 8 years 8.58 months, the critical ratio of the difference being 1.504, with chances of 93 in 100 that the difference was greater than zero.

The situation of the two groups was reversed for the Porteus Maze. Here the socially well-adjusted scored 8 years 8.16 months, while the maladjusted averaged 7 years 9.33 months, the critical ratio of the difference being 3.0. On an I.Q. basis, there was a difference in favor of the socially adjusted group of 8.62 points, which was again completely significant. Incidentally, the tendency to score higher in the Maze than the Binet, which was previously mentioned as lowering the incidence of mental defect, if Maze and Binet scores are combined, was again demonstrated. The Healy Picture Completion test showed only 2.15 points advantage for the well-adjusted, the difference not being statistically significant. This finding has a bearing on the inclusion of this test in the Wechsler-Bellevue Performance Scale. According to the above results, it has extremely doubtful validity as a test of social competency.

In 1932, three years later, Karpeles [10] repeated the study at the same hospital but used larger numbers of cases divided on the same basis as before. The average Maze test quotient of the socially-

adjusted (N = 185) was 94.3, that of the maladjusted 86, almost exactly the deficiency found by Poull and Montgomery.

When the cases were separated into two groups, above and below 80 I.Q.* by the Binet, the well-adjusted of superior Binet level averaged 100.5 T.Q. in the Maze. The high-level, maladjusted cases averaged only 89 points, or 11 points below the well-adjusted. Both differences were significant. Thus the Maze superiority associated with social competency was not dependent on intelligence level.

It is unfortunate that we still cannot adequately measure the degree of delinquency as a form of social incompetency. Through lack of such quantified data we have no correlation studies and are forced to rely on average differences in test scores between groups, which have been identified by their records as either socially well-adjusted or ill-adjusted. Though the comparative validity of mental tests is not exactly determined by cross-validation methods, it is interesting to note that the two independent studies cited above strongly support the claim that the Porteus Maze is a fairly valid index of social incompetence. But the warning should be given that the relationship is not necessarily close. As far as the Social Ratings Scale is concerned, coefficients between .6 and .7 do not support any claim for approximate correspondence or equivalence. However, tests deserve some latitude as regards confidence, since the criterion itself is far from having fully established validity. The matter of juvenile delinquency, as related to the Maze, will be discussed in a later chapter.

Chapter References

[1] Jacobsen, C. F., Wolf, J. B., and Jackson, T. A., "An Experimental Analysis of the Functions of the Frontal Association Areas in Primates," *J. Nerv. Ment. Dis.*, 82, 1–14, 1935.

[2] Porteus, S. D., "Mental Tests for the Feebleminded: A New Series," *J. Psycho-Asthenics*, 19, 200–13, 1915.

[3] Porteus, S. D., "Motor-Intellectual Tests for Mental Defectives," *J. Exper. Ped.*, 3, 127–35, 1915.

[4] Louttit, C. M., and Stackman, Harvey, "The Relationship Between Porteus Maze and Binet Test Performance," *J. Educ. Psychol.*, 27, 18–25, 1936.

[5] Stoddard, *loc. cit.*, p. 285.

[6] Wechsler, *loc. cit.*, p. 129.

[7] Porteus, S. D., *A Study of Personality of Defectives with a Social Ratings Scale* (Vineland, N.J., Research Dept. Publications, 1920), No. 23.

* The use of I.Q. in this book is merely following convention. T.Q. is the proper term.

[8] Porteus, S. D., *Studies in Mental Deviations* (Vineland, N.J., Research Dept. Publications, 1922).

[9] Poull, L. E., and Montgomery, R., "The Porteus Maze Tests as a Discriminative Measure in Delinquency," *J. Appl. Psychol.*, **13**, 145–151, 1929.

[10] Karpeles, Lotta M., "A Further Investigation of the Porteus Maze as a Discriminative Measure in Delinquency," *J. Appl. Psychol.*, **16**, 426–437, 1932.

CHAPTER IV

The Forebrain and Foresight

In validating any mental test three main questions call for attention. The first is, what does the test measure? If its relation to intelligent activity is clear, then the test is a good test in the sense that its further validation is worth while. The second question is to a large degree methodological. How well does it measure it? Like the first question, the third query relates to the usefulness or the meaning of the test results. How important to social adjustment is the trait or trait-complex that is measured? Upon this should depend its use in clinical practice.

As with all other mental measures, the answers to these questions in relation to the Porteus Maze cannot be more than partial or experimental. So long as this is the case, the evidence for the test's validity must remain to some degree presumptive. However, I believe that this evidence as regards the Maze is, in at least one important direction, extremely and almost uniquely impressive. It may be indirect or circumstantial in nature, but when direct evidence is not at hand, circumstantial proof is acceptable.

This evidence comes, not through the usual psychological channels, but from psychosurgery; that is to say, from brain operations undertaken for the relief or cure of psychotic or psychoneurotic symptoms. If the brain is the basis of mind and if its functioning as an organ of choice is related to the cortex of the frontal lobes, then damage to those lobes should result in impaired performance in a test specifically directed toward planning ability, the choice of alternative courses of action being the basis of planning. To support this hypothesis, we planned to apply the test prior to operation, then to repeat the measure as soon as the effects of surgical shock had subsided. Unfortunately for clear-cut demonstration, the Maze is subject to practice effects that differ with different individuals, with level of performance, and possibly with other factors. Hence, unless a practice-free or practice-controlled form of the test is used

postoperatively, the actual impairment in Maze performance may be diminished or concealed. In considering the results of earlier studies of the mental effects of frontal lobe operations, it should be kept in mind that the Maze Extension series was not then available, hence the size or gravity of deficits was partly screened by increases in scores due to practice.

It must also be remembered that both the nature of the test and the method of its application are intended to bring to light individual differences, which is its main value or purpose. For this reason we cannot expect postoperative losses to show equally in all individuals; we must therefore rely for test validation on group comparisons between control and operated patients. A working hypothesis is that if the Maze measures planfulness, then other tests which are not specially directed toward foresight should show lesser postoperative differences.

But before discussing this new approach to validation through neurosurgical operations, some attention should be paid to other interrelated questions. How well does the Maze measure planfulness, the trait complex it purports to measure, and how important is the trait measured? A third consideration is the degree to which the frontal regions mediate planning.

It must be recognized at the outset that the planning problem involved in the test is "situational," which poses the further important question whether ability, as shown by a high Maze score, passes over into real life situations where good planning is at a premium. Perhaps the answer would be plainer and more satisfactory if the question were asked in terms of ordinary rather than high level planfulness. A major portion of our discussion throughout this volume is devoted to proof that Maze Test abilities are related to everyday adjustments.

The position taken up in this book is that the Maze examines rather basic types of planning at a comparatively simple level. In other words, I doubt whether it is in any sense a measure of the level of foresight of the kind necessary in scientific and technological achievement. To use the Maze in this way would be analogous to testing higher mathematical ability by testing arithmetical skill. But it seems obvious that a mathematician could not possibly be adept in dealing with algebraic symbols if he lacked the ability to use simple number abstractions effectively, provided, of course, that he has had the interest and experience in arithmetic that he has had in algebra. To take another analogy, we may agree that a competent aviator need not necessarily be a skilled motorist, but I doubt very much if anyone could manage a plane very well if

after training and experience in handling a car he could not qualify as a driver. The greater ability may include the less, but the simpler ability does not presuppose the greater.

The belief that planfulness as shown in maze-threading is important would find support if it can be shown that injury to such an extensive and strategically placed part of the brain as the frontal lobes results in grave impairment of Maze test scores. Here again the argument is circumstantial. No one can tell from direct observation what work the frontal lobes are doing, but it is safe to assume that the outcome of such a long and involved evolutionary process as the building up of the human frontal lobes argues that they must have extremely important mental functions. What these functions are can best be determined psychologically by observing the behavior changes in operated patients. Unfortunately, no system of quantified data of social behavior is available. However, this matter must be discussed more fully later.

Factorial analysis by J. P. Guilford [1] seems to indicate clearly that there are two types of planning or foresight: perceptual and conceptual. Undoubtedly, the second type is the higher, but I find it impossible to make any clear-cut division between the two. In other words, in most activities involving planning, the two types of foresight are involved. Guilford reports his findings as follows:

> Porteus has maintained that his series of maze tests measure foresight. He can well claim support from the factor analysis just mentioned. The type of foresight measured by maze tests, however, is of a concrete variety. This ability may be important for the architect, the engineer and the industrial lay-out planner. It may not be found in the political strategist and the policy maker. So far as our results go, the maze test should by no means be offered as a test of general intelligence.

It is, of course, quite plain to anyone who is familiar with the situation that the Maze has never been offered as "a test of general intelligence" but rather as a test of a very important segment or area of general intelligence inadequately covered by any other series of tests. Guilford's suggestion that the Maze may examine the kind of concrete planning important to the architect, the engineer, or the industrial lay-out planner is, I am afraid, too flattering to the test. I would be quite satisfied with the claim that it tests the foresight of the man in the street, for whose ordinary adjustments the habit of careful preconsideration before action is essential.

As to the complex planning of the political strategist, the result of applying the Maze to Winston Churchill, for example, could be

expected to reveal nothing more than a supreme impatience with such trivialities.

The limitations of reference of the test have been repeatedly acknowledged by me.[2] One of my statements, published in 1939, is typical: "All that I claim for it is that it appears to be the best test available for testing prudence and planning capacity *at or about normal levels.*" What I have insisted upon, in and out of season, is that a measure of planfulness is most important and in difficult cases essential to mental diagnosis. In such cases, I believe that it should carry equal weight with the Binet or a verbal I.Q. obtained by the use of the Wechsler-Bellevue series.

Though Guilford's identification of two types of foresight, perceptual and conceptual, is undoubtedly well authenticated, it seems important to recognize that both factors affect successful maze-threading. There is an analogy between maze-threading and driving a car in heavy traffic. Obviously, perceptual anticipation is called for constantly in both activities. But along with that kind of mental alertness or vigilance there are conceptual factors not concretely represented. These include a vivid realization of the unhappy consequences of accidents, both to oneself and to others. In a prudent person this results in building up caution as against rash impulsiveness. Then, too, we are jealous of our social prestige. Most of us would think it an affront to dignity to be arrested for negligent driving. Not only is there embarrassment but there is also the cautionary factor of respect for law and order. Also there are positive considerations such as courtesy, respect for the rights of others and for property, our own and that of other people. Though in some of these motivations the parallel between maze-threading and driving a car cannot be maintained, nevertheless one concept that is fundamental to success in both operations is ordinary prudence, the kind of judgment that is expressed by the thought "I must take all reasonable care before committing myself to a course of action." This enters into so many other life situations that I believe it explains the wide reference and areas of validity of the Maze. The vigilance must be immediate, but at the back of it is the concept of constant caution.

Misapprehension of the task's difficulty often means a lowering of caution, which explains why highly educated people are very poor subjects for Maze Test research. When an adult sees that he is taking a test labeled "Year VIII" he frequently feels that the task is so simple that he ought to be able to do it with a minimum of effort. He may react to failure with impatience or contempt for the test, expressed by rejecting the task as too childish. He then may become inordinately reckless.

The combination of temperamental and mental factors in Maze performance undoubtedly diminishes its factorial purity, but this probably does not interfere with its clinical usefulness. It measures not only a person's ability but his tendency to put to practical use what ability he possesses. Penfield and Evans have suggested that what they call "planned initiative" is one of the functions of the frontal lobes, and this is essentially what I mean by planfulness or what Clement, a French psychologist, in a private communication, has called "utilizable intelligence." He has stated on the basis of studies with mental defectives, his belief "*Que votre test des labyrinthes est un bien meilleur predicteur de l'intelligence 'utilisable' que les tests verbaux, ou les tests factoriellement purs.*"

The theory suggested by Penfield and Evans [3] that planned initiative constitutes one of the main functions of the human frontal lobes is somewhat revolutionary as is my own suggestion that its testing should be an essential part of every assessment of general intelligence. The Penfield-Evans assumption displaces the notion that the frontal lobes are reserved for general intelligence. This earlier viewpoint was also supported by circumstantial evidence. A review of phylogenetic brain evolution is essentially the story of the rise of the neopallium, most apparent in the cortical development in the frontal lobes, and this evolutionary process kept pace, step by step, with the increase in comparative intelligence. An observer confronted with an array of the brains of organisms ranked according to their capacity to respond to a wide range of relevant stimuli could not help but be struck with the concomitant development in brain evolution, especially in the frontal region. The fact that this development is easily apparent was undoubtedly the reason for over-exaggerating its importance to intelligence. Without any special knowledge of brain functions, writers fell into the trap, so that the "noble brow" became, to their minds, the undoubted index of wisdom and, conversely, the low brow signified that its owner was among the stupid or inept. As a matter of fact, the great development of the other lobes of the brain was also a feature of human evolution, but it was perhaps not so obvious as regards gross anatomy.

Psychologists, in my opinion, have not realized the importance of the parietal, temporal, and occipital lobes to intellectual functioning. It is true that the growth of the frontal regions was novel and spectacular, whereas the development of the areas subserving perception was in the nature of an overgrowth of older phylogenetic structures, yet the vast increase in complexity and range of man's relevant stimuli made that overgrowth necessary. Inevitably, under the principle of neurobiotaxis, or the grouping together of re-

gions of related mental functions, one would expect that intelligence would not be entirely centered in the frontal lobes. One must not be led into any too literal analogy of a single commander or command post.

This exaggerated emphasis of the overlordship of the frontal lobes is well exemplified by Tilney, the famous neuroanatomist, who regarded them as the accumulators of experience, the directors of behavior, and the instigators of progress. If this were true, a man with extensive damage to his frontal lobes should exhibit extremely inferior intelligent behavior.

The support for such an expectation arose from the question, Why, if the frontal brain was not of such supreme importance, did Nature go to such pains to produce and house it? Why crowd out the smell brain? Why fissure the human brain so deeply and elaborately unless to add to our cortical equipment? Why lay down such a tremendously complex communication system, by which the frontal lobes were put in close relation to all other parts of the nervous system, unless they were the command posts, the organizing, correlating centers of human structure?

As previously mentioned, the evolution of distant sense receptors—smell, vision, and hearing—was necessary to the setting up of the anticipatory reactions of avoidance and attraction, making possible the decisions that are vital to the preservation of the individual and to the continuance of the species. These distant receptors paved the way through their development for the brain to become ultimately the organ of delayed choice. They made it possible to interpose reflective intervals between stimulus and response, a delay that was conducive to successful adaptation. Hence, there must be some coordinating center where the information from vision, hearing, touch, and smell could be correlated and classified in the light of past experience before the motor centers could be innervated and appropriate action anticipated.

In suggesting the frontal lobes—including the cingulate gyrus and, following McLean's proposal, the upper brain stem—as the coordinating center where decisions are planned and determined, there is no intention of overlooking the contribution of the other great cerebral divisions to human survival. The specific intellectual abilities, which are not so directly essential to existence, may be mediated elsewhere. It may well be that cerebral position rather than cerebral extent is the true index of survival value. Certainly Nature will not put all her eggs in one basket, even as convenient a basket as the frontal lobes.

The fact that animals rather low in the evolutionary scale must

also possess mental alertness to a high degree and some measure of foresight in order to survive indicates that these functions are mediated also by brain regions other than the frontal lobes. According to Fulton,[4] these latter are late in being differentiated. He says.

> The frontal lobes emerge as a recognizable division of the cerebral cortex in all vertebrate forms but the general pattern found in the human brain occurs only in the primates. Even in lowly tarsiers and galagos the rudiments of a central sulcus is present in the form of an inconspicuous dimple in an otherwise smooth and virtually unconvoluted forebrain. The major divisions of the frontal lobe itself do not become clearly defined, however, in tarsiers or marmosets, but they begin to appear in the higher lemurs and are sharply defined in the old and new world monkeys.

With such evidence from comparative anatomy that the forebrain constitutes a steady evolutionary development, culminating in the human frontal lobes, it was no wonder that the assumption was made that it was also the principal or primary seat of intelligence. In this matter, psychologists followed the suggestions of the neurologists.

But in the last 20 years neurosurgeons have carried out daring operations on this region, effecting the most extensive and deep-seated damage. Their surgical procedures showed that the brain was anything but "the soul's frail dwelling house," as Shakespeare termed it, but a really tough organ. Lobotomy, carried out by various methods, severed a very large proportion of the nerve pathways connecting the frontal cortex with the thalamus and hypothalamus. Topectomy, involving actual excision of whole areas of cortical brain substance, can be almost equally drastic, while "scrambling" the patient's brain in the anterior region by transorbital lobotomy certainly constitutes what may well be called a grave surgical "insult." If, then, the frontal lobes alone really constitute the seat of intelligence, such extensive brain damage could be expected to render the postoperative subject mentally incompetent and little better than a zombie. But, most surprisingly, mental examinations showed clearly that while in some cases there were impairments in general intelligence, on the average this expectation was very far from being fulfilled.

It is no wonder that psychologists, such as Carney Landis, were extremely impressed with the dearth of evidence relating the forebrain to intelligence as it is commonly defined. No one has had greater experience in this matter of the mental effects of psychosurgery than Landis.[5] He stated in 1951:

This portion of the brain has been held to be the seat of highest intellectual ability; the lobes have been called the greatest achievement in evolutionary progress; the frontal cortex has been said to be the measure of intellectual development of the human race, and so on and so on. One may speak of this era as that of "the reign of the myth of the frontal lobes."

With respect to the mental effects of frontal brain damage, Landis makes an exception in the case of more posterior operations by stating: "There is fairly good evidence that the nearer the surgeon gets to the primary motor areas during the operation, the more apt one is to find postoperative effects which are undesirable."

Of postoperative changes, Landis remarks:

In most, but not all, patients there is a decrease in mental test performance which was shown best by the Porteus Maze Test. . . . This test showed rather regularly a drop of one or two years in mental age, which was regained during the period of three to six months after operation. No other intelligence test shows that much of a deficit, and those that did show minor losses during the first three weeks after operation, showed the loss to be regained completely three months postoperatively.

He asks: "Is this a temporary lack of foresight?" A reappraisal of this situation will be presented later.

Wilder Penfield,[6] the well-known neurologist, held a much more positive view. He said: "The large sheet of cortex which covers the front part of the brain seems to be utilized by man when he is thinking of new plans and seeking greater insight into life's problems." Apparently this opinion was largely based on his observation of the effects of lobotomy, for he says: "The luckless individual does forfeit capacity for planned initiative. He may have gained peace of mind but he has lost a type of insight that is difficult to define." Though insight may be not only difficult but almost impossible to measure, it is possible to obtain a fairly reliable index of planned initiative. This disagreement between psychologist Landis and neurologist Penfield probably depends on different viewpoints as to what are to be regarded as higher mental functions.

My own view is in close accord with Penfield's in so far as I am unable to consider planning apart from intelligence. To my mind, the unplanned action has either gone beyond the learning stage and has become automatic, or it is instinctual in nature.

In 1929, before the advent of psychosurgery, Professor Wood Jones and I [7] set forth our ideas of the functions of the frontal lobes in a book called *The Matrix of the Mind.* We wrote:

In the normal human being the answer to that ever-insistent question—What shall I do?—is provided after an interval appropriate to the circumstances. In human beings departing only slightly from the normal, there may be more or less than the deliberation that suffices for the truly normal. An individual may be normally prudent, more than normally prudent, or less than normally prudent, as we would say in reviewing the conduct of our fellows. . . . Again at the other extreme are those who act entirely on impulse on all occasions, and for whom a considered line of ordinary social conduct is quite unattainable. In lesions of the frontal area this is the type of conduct produced.

Awareness then has led to this: that it has given us a typically human attribute of the considered actions of the individual towards his human environment. It has given us what we know as human social adjustments of conduct; and the special pallial seat of this human attribute is in that typically human region of the pallium known as the frontal areas.

I believe the main factor responsible for the swing toward skepticism regarding the theory that the frontal lobes have much to do with high mental functions is the acceptance of a too-restricted definition of intelligence, which omits any reference to planfulness. Another consideration has been referred to previously—the paucity of tests of foresight. Had there been as many tests available of this function as there are of so-called general intelligence, the weight of evidence for postoperative change would have been far more impressive. A third hindrance to demonstrating the extent of deficits in planfulness following brain damage was disregarding the fact that the Maze is particularly susceptible to practice effects. Hence, a regain of deficits in test scores did not justify concluding that the loss of foresight was merely transient, since the practice effects of reapplying the same form of the test merely canceled or concealed the more permanent losses. Fortunately, new evidence on this last point is now available and will be cited later.

If this position is accepted, we are no longer in the quandary of having to find functional reasons for the development of the forebrain. The lobes themselves are no myth—even though ascribing sweeping powers or mediation of general intelligence as their function may need modification. No comparative anatomist would accept for a moment the idea that the forebrain is an evolutionary luxury. It seems to me that common sense demands that we consider planning and foresight as fundamental elements in mental life. The experimental evidence that their exercise is related to the frontal lobes will be presented in the next chapter.

I am by no means alone in believing that frontal lobe damage results in impairment of planning capacity. Clinical observations

have led to similar conclusions. Some of the most relevant of these are presented in the next chapter, but at present we shall quote some typical statements taken from various statements by Freeman and Watts, the pioneers of psychosurgery in America:

> Lack of foresight seems to be one of the most characteristic manifestations of deficiency in frontal lobe activities.
>
> Patients lack the drive to pursue a persistent yet flexible course towards a distant goal.
>
> Truly, if there is any way of summing up in one word the function of the frontal association areas, that one word should be foresight.
>
> Substantially, the frontal lobes, in our opinion, add foresight and insight. . . . Foresight is badly impaired for a period after operation. . . . Creative work of any considerable complexity is a closed book for most patients with frontal lobotomy, although there are a few shining examples of superior performance.

They then raise the question whether the definition of intelligence should be "broadened to include planning capacity and enterprise, the setting of a goal, and planning the course to attain it, and the adaptability to the occupation." They regard this as an "unwarranted extension" of the term because "so many other factors, particularly emotional ones, enter into the total reaction of this type." [8] The position taken up in this book is that we cannot avoid including planfulness as part of the definition of intelligence since no intelligent activity can be effective without it. Admittedly, emotional factors can interfere with initiative and hence with intelligent planning, but surely it is only a matter of descriptive convenience to talk of emotion, planning, initiative, and intelligence as if they were distinctly separable components of an intelligent act. Certainly, in Maze performance they all interact.

Recent neurological discoveries throw new light on emotional factors as they are related to frontal lobe functions. An anatomical complex, formerly part of the rhinencephalon or smell brain, has been identified as serving autonomic and emotional factors. MacLean [9] has called this older portion of the central nervous system "the visceral brain." Its interconnections with the prefrontal cortex and with the hypothalamus and the upper-brain stem bring its functions within the orbit of intelligence. MacLean makes a psychological distinction between feeling and knowing, but between the two there is, in my view, an inevitable interdependence. What and how we feel is affected by what we know or believe. Conversely, emotional factors often determine what we accept or reject as knowledge.

The situation would, I believe, be clearer if we can accept the notion that the upper-brain stem, with its more immediate frontal connections, constitutes what might be considered the nearest approach to a command post where decisions are made and executed.

Fulton [10] deals with this matter in the following statement:

> MacLean draws the conclusion that the phylogenetically old brain is largely concerned with visceral and emotional reactions—a region which "appears to be so strategically situated as to be able to correlate every form of internal and external perception. . . ." Intellectual functions, on the other hand, are associated with the highest and most highly developed parts of the brain, and again to quote MacLean: "This situation provides a clue to understanding the difference between what we feel and what we know."

However, Fulton's statement with regard to the *locus* of intellectual functions should be read in connection with the recorded results of psychosurgery, unless he considers planfulness an intellectual function.

He also quotes Ward on the effects on monkeys of operation on the cingulate gyrus. Ward noted that the animals seem to have greatly diminished fear and anxiety "and thus lose the ability to forecast the social repercussions of their own actions." At their level it is evident that planning is related to the functions of the cingulate area; the animals no longer know what to fear. It is noteworthy that a standard lobotomy involves the cingulate gyrus in humans.

Landis [11] and his co-workers emphasize the disappearance of morbid anxiety, which may rise to the pitch of anguish in psychotics, as one of the outstanding benefits of psychosurgery.

Robinson and Freeman [12] also are convinced that at the roots of the mental troubles cured or improved by lobotomy is "a morbidly elaborated sense of self-continuity, which involves a chronic looking backward and forward."

It would seem clear that both observation and experiment agree in showing that a relationship exists between morbid anxiety and heightened anticipatory reactions. Some people are oversensitive to possible threats to life or happiness, are always crossing bridges before they come to them, or, as we say, meeting troubles half-way. When this apprehension reaches abnormal heights, the burden of fear and imagined terrors may become so overwhelming that it stultifies normal resistance. The unfortunate patient may then erect all kinds of bizarre defenses, such as delusions of grandeur or the fantastic localization of sources of danger as in a persecutory com-

plex. If through an operation the neural mechanisms of anticipation and prehearsal suffer a structural and functional reduction, anxiety is allayed. In short, if vigilance is blunted, threats, either real or imaginary, lose their emotional affect. An unrecognized danger does not call for defense mechanisms or a rationale based on delusional ideas. In the case of intractable pain, the dulling of emotional reactivity makes it bearable.

But this is the abnormal picture. Under normal conditions vigilance is a fundamental factor in human survival and anticipatory projections are essential to thought and invention. The Maze examines this capacity at a rather simple but basic level. The functions it tests find representation in both the oldest and the newest developments of the central nervous system.

For the concept of vigilance I am indebted to Landis and his co-workers. He follows Sir Henry Head in his description of this most important trait as being indicated by "signs of integration and purposive adaptation." On the other hand, diminished drive and initiative after psychosurgery have been noted by observers such as Arnot and Partridge. Landis and Zubin [13] confirm this lack of drive and energy. They say that in the early postoperative period "it was as if the patient were sleepy, fatigued or partially drunk so that there was disinclination to work at his usual level." However, my own experience with the Maze did not reveal any such "disinclination." The reactions seemed to me to be a lessened awareness of blind alleys as traps, an obtuseness often shown by the repetition of identical errors in successive applications of the same Maze design. Patients apparently tried just as hard but were certainly less concerned over their mistakes than formerly. Frustration was by no means as disturbing.

Though not part of the forebrain, the reticular formation in the brain stem seems to be related to wakefulness and mental alertness. If further physiological observations confirm this view, then its functioning would also affect maze-threading, which would further support my theory that the Maze reflects mental states and abilities related to survival.

Maze deficits are most marked, as we shall see later, in injuries to areas 10, 9, 8, 6, 46, the cingulate area 24, and to portions of the temporal lobe. These injuries also interfere with the neural pathways of the above cortical areas connecting them with thalamus, hypothalamus, and reticular formations. These regions constitute two divisions of what Fulton calls "the orbito-temporo-frontal cortex." Topectomy of orbital regions 11 and 12 and transorbital lobotomy have less effect on Maze performance. The inference is

strong, therefore, that the Maze stands apart from the other more purely intellectual tests because of its psychosomatic relationships.

Chapter References

[1] Guilford, J. P., "The Structure of Intellect," *Psychol. Bull.*, **55**, 267–93, 1956.

[2] Porteus, S. D., "The Validity of the Porteus Maze Test," *J. Educ. Psychol.*, **30**, 172–78, 1939.

[3] Penfield, W. G., and Evans, J., "The Frontal Lobes in Man: A Clinical Study of Maximum Removals," *Brain*, **58**, 115–33, 1935.

[4] Fulton, John F., *Functional Localization in Relation to Frontal Lobotomy* (London, Oxford University Press, 1949).

[5] Landis, Carney, "Psychological Observations on Psychosurgical Cases," *Psychiat. Quart.*, 1–9, July 1951.

[6] Penfield, Wilder, *The Physical Basis of Mind* (Oxford, Blackwell, 1950), Chap. VII.

[7] Jones and Porteus, *loc. cit.*

[8] Freeman, Walter, and Watts, James W., 1942. *Psychosurgery* (1st ed., Springfield, Thomas, 1942), p. 154.

[9] MacLean, P. D., "Psychosomatic Disease and the 'Visceral Brain,'" *Psychosom. Med.*, **11**, 338–53, 1949.

[10] Fulton, John F., *Frontal Lobotomy and Affective Behavior* (New York, W. W. Norton, 1951).

[11] Landis, Carney, "The Frontal Lobes and Anguish," *J. Nerv. and Ment. Dis.*, **115**, 203–14, 1952.

[12] Robinson, Mary F., and Freeman, Walter, *Psychosurgery and the Self* (New York, Grune and Stratton, 1954).

[13] Landis, Carney, and Zubin, Joseph, *Psychosurgical Problems* (Philadelphia, Blakiston Co., 1952), Chap. 14.

CHAPTER V

Psychometry and Psychosurgery

The advent of tranquilizing drugs as a less drastic approach to the relief of psychotic symptoms than psychosurgery, in many cases with similar effects, has resulted in a marked decrease in the number of lobotomies, topectomies, and other forms of prefrontal surgery. Except in special cases of intractable pain, psychomotor epilepsy, etc., psychosurgery seems to be falling rapidly into disuse. It may therefore seem strange to some readers that so much attention to its results in relation to the validation of psychometric instruments should be accorded to this topic in this chapter. But the fact that the operative procedure bids fair to be outmoded is all the more reason that these results should be carefully summarized and evaluated. In my opinion the psychologists' neglect to engage in studies of lobotomized or topectomized patients while they are still available represents a lost opportunity. "For the first time in history," says Hutton,[1] "an opportunity has been presented for the study of changes in personality produced by a relatively standardized local lesion of the brain." Yet the amount of recorded research is not large. Considering the difficulty experienced in setting up any external criteria for the validity of mental tests, the comparative indifference of psychologists is astonishing.

Dr. Thelma Hunt[2] was apparently the first psychologist in this country to attempt measuring the effects of lobotomy by mental tests. Her results were published in the first edition of the book by Freeman and Watts, which reported their own pioneer work in psychosurgery. Several years earlier, Rylander[3] had made a very careful and complete report on personality changes following unilateral frontal lobectomy, but they may differ from the results of the bilateral operations.

Hunt applied a variety of tests to forty of Freeman's and Watts's lobotomy patients with what may be called negative or indifferent changes apparent as the result of operation. The Kohs Block

Design test, widely recognized as a valuable clinical instrument, showed 33 per cent of the patients improved, 37 per cent had lower scores, and 30 per cent remained the same. The tests that showed statistically reliable differences were speed in Cube Construction, Arithmetic, Color Naming, accuracy in cancelation, together with scores in the Match-Stick and Coin-under-Cups Test, but all of these were in the direction of improvement after operation. In a Paper Form Board Test, patients were a little slower on the average but considerably more accurate. Speed in the Substitution test was reduced. Hunt summed up the situation as follows: "Taking these tests together or as a whole, there is no evidence of appreciable change in general mental ability."

Hunt's study of Rorschach responses also failed to reveal any significant changes, particularly those that could be considered quantitatively. Qualitatively, she noted fewer neurotic trends, and, in general, "a more restricted and constricted personality," and less variety and individuality of response than are found in normal subjects. Inhibition of spontaneity and creativeness were also indicated. If, as is supposed, the Rorschach reflects personality, it is surprising to find that lobotomy, which so often brings about such remarkable reconstructions of the patient's personality, affects to such a small degree Rorschach reactions. However, this is no more surprising than to discover that, even in patients whose ultimate social adjustment is good, there should be a severe decline from their initial Maze scores.

With regard to Hunt's finding of no appreciable change in general mental ability, she has adequate support from other investigations. The most striking confirmation of her results comes from the Columbia-Greystone Projects in the field of prefrontal topectomy, a different surgical approach, but one which, like lobotomy, constitutes grave damage to the integrity of the frontal lobes. These projects will be discussed later. Our present purpose is to discover what has been changed in the average patient and whether there are deficits that can be related to Maze-tested mental functions.

Unfortunately, the personality changes are of so diffused and general a nature that, though obvious to careful and intimate observation, they can only be vaguely described. For example, in the period up to six weeks after operation, many patients can best be characterized as childish. This is shown by their simple interests and amusements, by their need for prompting to carry out ordinary personal activities, such as bathing, going to bed, getting up in the morning. They are also characterized as listless, easily

distracted from executing simple jobs. Nevertheless, they are cooperative, often persisting in repetitive actions such as folding a sheet over and over again, or polishing the same window pane. Freeman and Watts characterize this type of conduct as "surgically induced childhood" and state that the patient "regresses to an infantile level of behavior." Typical reports by ward attendants usually mention laziness, apathy, somnolence, inaction or restlessness, repetitive activities, psychomotor slowing. One of the most common comments cites lack of initiative and planning. These defects in many cases lessen or disappear with time. Hence, it is evident that the operation's result can be expressed in terms of both profit and loss. Apathy, loss of initiative, and emotional blunting may make the social behavior more tolerable to others, but these gains in sociability also conceal other losses, such as diminished planfulness.

The realization that personality has been adversely affected without the observer being able to define the changes accurately is illustrated in a most thoughtful discussion of the subject by Partridge,[4] who studied 300 postoperative cases in England. Speaking of 60 patients considered to have recovered from affective disorders, he stated that nearly all showed demonstrable changes with undoubtedly undesirable aspects after prefrontal operations. He summarizes these trends as lowered drive, reduced activity and initiative, less responsiveness, as well as diminished intensity of emotional experience. Partridge called the typical patient more "primitive," with less awareness, resulting in a less critical attitude both as regards himself and other people, and an avoidance of any intense efforts. But he notes that if these characteristics are found before operation, they are more likely to get worse.

In the light of continued experience, I would express the same notions in different terms by saying that with less clear-cut anticipatory reactions or prehearsal, a patient's decisiveness in social behavior is decidedly hampered, which is also reflected in his impaired ability to make the type of decisions called for in the Maze. This view is supported by Freeman and Watts, who cite "limited capacity for conservative and constructive thought, lack of deliberation or weighing of possible consequences, badly impaired foresight, failure of drive to pursue a persistent yet flexible course." Ström-Olsen, Last, and Brody[5] also remark on loss of initiative, spontaneity, lethargy, and vacillation in their patients. Hutton also mentions incapacity for forethought, so that "the sufferer is not himself. His best qualities are in abeyance, revealing the second best. He is slightly less of a man, and more of a child." With re-

gard to patients in general: "They are not incapable of taking thought for tomorrow should anything stimulate them to do so, but they are content for tomorrow to take care of itself. They live in a perpetual present that is dependent upon associative memory."

Partridge also comments on defects in routine planning owing to lessened deliberation. He suggests that "lack of restraint facilitates the too hasty arrival at decisions in that associational arrangement of affairs that we know as planning, and thus contributes to an inadequate performance in conceptual thinking, which we recognize under the term lack of judgment." This viewpoint coincides with the statement of frontal lobe functions, previously quoted from Wood Jones and Porteus.

On the other hand, observers seem to agree that the commonly used mental tests are mainly insensitive to these changes. With the exception of the possible impairment of categorical and abstract thinking, psychosurgery seems to have occasioned little retrogressive change according to such investigators as Kisker, Halstead, Carmichael and Bucy, and Robinson. However, Partridge sums up his studies by stating that he cannot accept the view that intelligence is unaffected by operation. "It would appear," he says, "that the tests at present used are not sufficiently refined." All these pronouncements of no significant change after operation serve only to throw into bolder relief the impairment that takes place in most, though not all, psychosurgical patients in planfulness as measured by the Porteus Maze.

In 1942, when Freeman and Watts published their book *Psychosurgery,* I began studying the lobotomy cases that had been operated upon by Dr. Ralph Cloward, noted neurosurgeon, at the Kaneohe Territorial Hospital. The results were published in collaboration with Dr. Richard Kepner,[6] then clinical director of the hospital.

Seventeen patients were available for testing, both before and after operation. It should be stated at the outset that the Maze was applied with the confident expectation that scores would improve postoperatively to correspond with the personality changes that made the patients easier to control and live with. But in spite of practice effects due to a reapplication of the same form of Maze Test, only two patients showed any increase in score. One of them then suffered a more posterior operative cut, and his score at once fell 11 T.Q. points below his preoperative level. To serve as controls, the records of 17 cases who had been examined twice by the Maze were taken at random from the Psychological Clinic files. Of these, 71 per cent had gained in the second examination as

against the 6 per cent of lobotomy patients who had improved. The net *gain* of the unoperated defective cases was 1.94 years, whereas the net *loss* of the lobotomy cases was 1.97 years, or about 14 T.Q. points.

In contrast with this deficit, the operated patients had a mean gain of 3 Binet T.Q. points, although 5 out of 18 showed significant losses. It was perhaps significant that all of the five had high Binet scores preoperatively—in other words, they had more to lose. As regards the Maze results, it seemed evident that while the type of social adjustment marked by more passive acceptance of environmental circumstances was not reflected in diminished scores, the lessening of vigilance, active planning, and initiative certainly were.

Immediately after the publication of this study in 1944, Dr. Henry Peters and I [7] began a more comprehensive investigation of Maze reactions in 55 lobotomy cases. The preoperative mean score was 11.29 years (S.D., .35), the postoperative 9.65 years (S.D., .39), or a mean loss of 1.64 years. This was equivalent to 11 points of T.Q. level. Notwithstanding practice effects, it was not until the fourth postoperative examination that the average score reached the preoperative mean.

As controls we used a group of 55 prison inmates who had scored 12 years on a first Maze examination and during the course of time had been retested. The group as a whole was of the same social grade as the patients and had a similar racial background. It would have been better in some ways to have had nonoperated hospital patients as controls, but this research design, too, has unsatisfactory features. The experimental group had undergone treatment, which in many cases had effected temperamental changes that might be expected to affect Maze responses favorably; any other hospital group had not. In my opinion, the problem of arranging controls among any abnormal population is a complex and difficult one.

The second application of the test to the prisoners took place after an average interval of time of 4.7 years, yet practice effects, undoubtedly in mental set, were plain. Only 14.5 per cent had lower scores on the second test, as against 65.5 per cent of the lobotomized patients. Only 27 per cent of the latter group gained, as against 69 per cent of the prisoners. The average elapsed time between testings of the lobotomy cases was 7.7 weeks, so that practice effects could be expected to be greater than after an interval of 4.7 years. The gain for the prison group was 1.6 years, and if this is the normal amount of practice effect for cases averaging 12 years on the first application, the actual change due to operation

would be 1.6 plus 1.65 years, or an actual deficiency of 3.25 years. This would be equivalent to 23 T.Q. points.

Patients were then divided by the hospital psychiatrists, according to their degree of social improvement, into three subgroups, unimproved, improved, and much improved. The criterion for the third subdivision was successful parole or discharge as socially competent persons. As a whole, this group was characterized by initial marked deficits in Maze score followed by steady or consistent increases on successive examinations until a point close to the ceiling of the test was reached. When charted, the graph looked like a root sign or a check mark ($\surd$). This was the pattern found also for the most improved cases in the Columbia-Greystone study, to be described later.

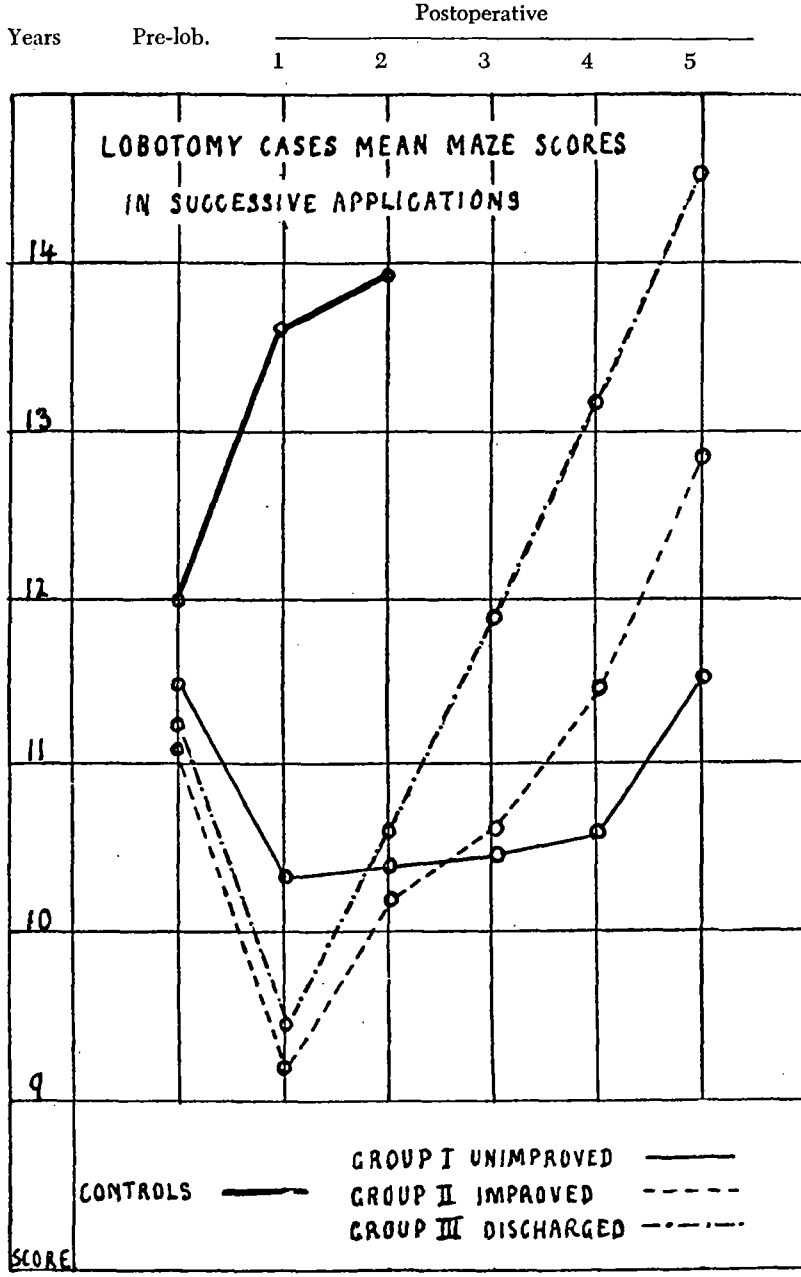
The improved group showed a rather similar pattern except that its regain in score on repeated tests showed a slower or postponed practice effect, the angle of the check mark being wider. The preoperative level was reached in 4 retests instead of 3 for the greatly improved patients. The unimproved group had a lesser initial loss, and the preoperative level was not reached until the fifth retest. The graph showing the above results is included in this chapter.

The above results have been discussed in earlier publications, and the picture seems to be sufficiently clear. However, there was one feature, not greatly apparent at the time, that has lately become of great significance. This is the somewhat puzzling fact that some patients' worst deficits did not show up at the first postoperative but at some later retest. While 66 per cent of lobotomized cases declined in their first postoperative Maze, the percentage of those showing, at some subsequent examination, a score lower than their preoperative level rose to 82 per cent. The percentage of patients declining in Maze scores agreed with that reported by Landis at the third research conference on psychosurgery in 1951.[8] He stated:

> Dr. Porteus had previously reported this sort of loss after lobotomy. We confirmed his finding that a brain operation performed on the frontal lobes gives rise to an immediate postoperative loss in mental age of 1 to 2 years, in some 80 per cent of psychosurgery cases. None of the other tests of intelligence or of intellectual function given in the Columbia-Greystone Project I gave any systematic evidence of interference with efficiency of performance, either scored in the form of a weighted test, scored as subtest, or scored in any of a dozen ways we devised.

In the ensuing discussion Hebb specifically excepted Robinson's and Porteus' studies from his general statement that there was no

FIG. 1. MENTAL CHANGES AFTER LOBOTOMY.



(Reproduced from Porteus, *The Porteus Maze Test and Intelligence* (Pacific Books, Palo Alto, Calif., 1950), p. 84.)

evidence of relationship between the frontal lobes and intelligence.

The matter of later postoperative declines had been discussed earlier very briefly by Peters and myself. We wrote: "Supported in many instances by ward observations is the theory that previous mental tendencies tend to persist for some time after the operation, running on their own momentum, so to speak, so that the full effects of lobotomy are not apparent for some time." If, as seems very probable, mental efficiency depends in part on blood chemistry as well as on localized brain functions, it would be feasible to suppose that the full effects of brain injury are not at first apparent. Furthermore, even in carefully circumscribed surgery, degenerative changes are likely to spread to neighboring anatomical areas, so that the full mental effects are deferred.

Aaron Smith's [9] findings in a study of Maze performance almost nine years after operation have a critical bearing on this point and will be discussed later. The suggestion that conservative habits of action that are deeply imprinted on personality may continue after operation, even when the frontal areas responsible for their initial acquirement are removed or damaged, seems to be quite reasonable. If true, it would account for the early successes followed by later impairment.

Another feature of Maze impairment, frequently but not consistently noted, is the subject's good performance in the easier designs followed by his rather sudden failure to solve the more complex mazes. Such an onset of mental confusion is often associated with repetitive errors. This "shatter effect" was first commented upon by Brundage * in a study of simple industrial aptitude, who noted it particularly in factory girls with high test records in general intelligence tests but with low performance in industrial output. The tendency to repeat the same errors seems to be accentuated after lobotomy.

Petrie [10] in England reported on 20 patients who were examined preoperatively and then postoperatively on two occasions, the first testing three months and the second nine months after operation. The subjects were not psychotics, "the vast majority of them being severe neurotics." Petrie also noted the same difficulties in using an adequately comparable control group that I have experienced. She found a mean loss of 7 points in Wechsler Verbal T.Q.s, a finding somewhat contrary to experience in this country, and possibly a result of the high intellectual level of her group. Significant declines were also apparent in the Full Scale I.Q. and in

* Private communication, December 26, 1943.

Comprehension. The only other significant loss was in the Porteus Maze, which amounted to 13 T.Q. points. It must be remembered that the form of the test used in the two postoperative examinations was the same as that used in the preoperative testing. This, of course, partially vitiates the force of her findings, which were that a significant loss occurred three months after posterior (standard) lobotomy, but that the loss nine months after operation did not quite reach the level of significance. However, Petrie suggests that "the absence of a significant loss after the longer interval is probably due to the increased practice effect which is known to be pronounced on this particular test; that is, the patients' deficit is somewhat hidden by their having had two previous trials at these same mazes."

Petrie also found, as I did, a significant increase in the number of double (repetitive) mistakes made by the group in both postoperative testings. She also noted that the speed of execution was greater after lobotomy. Both of these findings are fully confirmed by a recent study of neurotics to be discussed later. The increased speed of execution agrees with the Freeman-Watts clinical observation that there is less deliberation or restraint in lobotomy patients. Petrie's finding was the more striking because the postoperative records showed many more mistakes, and at each error the test must be begun over again, a time-consuming process. The greater speed record is the more remarkable.

Another of Petrie's findings confirms a fact previously found by others, viz., that damages inflicted on more anterior or rostral portions of the frontal lobe result in much less Maze impairment than if they are inflicted at a more posterior site.

Overshadowing all other investigations on the effects of psychosurgery are the Columbia-Greystone Projects I and II and the New York Brain Study. In research design, extensive collaboration by authorities in all the interrelated fields concerned, in full and competent reporting and discussion, these studies are outstanding. The results have been presented in three volumes, while the psychological findings have been the subject of several articles in the appropriate journals. For these reasons, only a very brief discussion of these projects will be attempted here, and that only in relation to the Porteus Maze and some comparative results with other tests.

The reason for including the Maze among the extensive battery of tests, yielding 135 different scores, is stated by King[11] in his presentation of changes in intellectual function following topectomy, or excision of cortical areas in the frontal lobes. After recit-

ing my views on the inadequacy of Binet-type psychometric examinations, which have been validated against educational rather than social criteria, King stated: "Since specific claims have been made that this test measures planning, initiative, and common sense judgment in practical situations, it would be expected to reflect alterations of these traits of intelligence in postlobotomy cases if, indeed, these traits are as modified as the accounts of clinical observation would lead us to believe."

The deficits shown in Project I seem unequivocal. The operated group had a mean loss in Maze score of 1.21 years postoperatively; the practice effects in the second testing with controls were marked, and these did not appear in the operated group who were below their preoperative level. Failure at the upper age levels was apparently responsible for this decline. Intrusion upon the superior area (nearer the motor area) caused the greatest decline, while damage to Area 11 (inferior orbital region) caused the least.

With regard to the Wechsler-Bellevue scale, only two subtests of the Performance Scale, the Object Assembly and the Digit-Symbol Substitution, showed any significant difference, mainly in diminished ability to profit by practice, but in both tests the operated group gained in raw scores over its preoperative level. Actually, in only one test (Similarities) was there a slight fall in raw scores after operation. This bears out the view previously mentioned that general intelligence scales do not show any consistent declines after extensive damage to the frontal lobes.

This brings up one minor point of criticism. The Wechsler-Bellevue results are all reported in T.Q.s; the Maze, in years of test age. This may hinder the reader in making adequate comparisons. In Table 6 the mental ages, as given by King in his Tables 43 and

TABLE 6

TEST SCORES BEFORE AND AFTER PSYCHOSURGERY
COLUMBIA-GREYSTONE I DATA

Test	Preoperative T.Q.	Postoperative T.Q.	Difference
Wechsler-Bellevue Full	101.7	105.6	+3.9
W-B Verbal *	112	114	+2.0
W-B Performance *	87	92	+5.0
Porteus Maze †	94.3	85.7	−8.6
Controls	94.3	101.4	+7.1

* Wechsler-Bellevue T.Q.s taken from Wechsler I.Q. table, 17–19 years.
† Porteus Maze T.Q. based on adult average age 14 years.

47, have been converted into T.Q.s so that the actual differences will be apparent. Some critics have questioned the adequate selection of controls. From what has been previously said, it should be clear that I also am critical of all control checks when subjects are drawn from a mentally abnormal population. In the table, the control group's Maze scores are also quoted to indicate that the 8.6 point loss for operated patients was in contrast with the practice gain of 7.1 points in the controls.

My major objection to the Columbia-Greystone conclusions is that while practice effects in the Maze were not only acknowledged but specifically demonstrated in the research design, their occurrence and importance seemed to have been overlooked in the statement of conclusions. Over and over again reference was made to the transient nature of the patient's Maze Test losses when it seemed obvious that his apparent recovery was merely a practice effect and not an actual regain of function impairment. It would have been better merely to report the losses without any reference to their duration or permanency. There was no conclusive evidence on this point, for or against, and it is only lately that the permanency of defects has been demonstrated. It is unfortunate that the practice-free Maze Test Extension was not available until 1955. It is strange that in other studies on various problems undertaken since that time there has seemed to be little disposition to control practice gains by the use of this form of the tests.

The second Columbia-Greystone Project was mainly notable for its investigation of surgical approaches other than ablations of the cortex. Venous ligation of two types—the first of the vessels draining the brain surface in a more anterior region of the frontal cortex, the other of an area extending 5 cm. caudal or more posterior—was carried through. These operations restricting cortical blood circulation interfered with the efficiency of the affected areas just as surely as a topectomy would have. The anterior ligation was followed by an insignificant gain in Wechsler T.Q. and a similar inconsequential loss in Maze T.Q. But whereas more posterior ligation made scarcely any change in Wechsler-Bellevue T.Q.s, there was a loss of no less than 28 T.Q. points in the Maze. This would indicate that the superior prefrontal cortex does subserve planfulness but not intellectual abilities as measured by the Wechsler.

Transorbital lobotomy, another anterior operation, resulted in a *gain* of about 3 points (to the nearest whole number) in the Wechsler-Bellevue, a *loss* of 6 points in the Maze. Thermocoagulation of areas 9 and 10 resulted in a *gain* of about 9 Wechsler-Bellevue I.Q. points, a *loss* of 11 Maze T.Q. points. Thalamotomy,

which destroys one of the terminals of cortico-thalamic communication, brought about a relatively tremendous deficit in the Maze of 30 T.Q. points, but a slight gain (2 points) in Wechsler-Bellevue performance. Thus the closer the injury comes to the central highway of communication between the superior cortico-thalamic areas, the cingulate region, and the hypothalamus and possibly the reticular formation, the more serious the Maze loss. The figures are shown in Table 7.

TABLE 7

I.Q. GAINS OR LOSSES AFTER PSYCHOSURGERY
COLUMBIA-GREYSTONE II DATA

Operation	W.-B. Verbal *	W.-B. Perf.*	W.-B. Full *	Porteus Maze †
Extensive Venous Ligation	−5.75	+8.0	+0.75	−28.0
Limited Venous Ligation	+2.43	+2.15	+1.48	−1.5
Transorbital Lobotomy	+1.86	+2.58	+2.86	−5.7
Thermocoagulation	+5.50	+13.0	+9.50	−10.7
Thalamotomy	+7.5	+4.5	+2.0	−30.0

* Wechsler-Bellevue results from King's Table 6, *Psychosurgical Problems.*
† Porteus Maze T.Q.s based on King's Table 14, *loc. cit.*

Unfortunately, the variety of operations diminished the number of patients in each category, but there is no solid ground for criticism in this respect. When students of the subject of brain localization consider the tremendous difficulties in the way of human experimental evidence, they are happy to learn of the facts regarding the results of thalamotomy, even on two subjects. The amount of deficit cannot be determined exactly by such data, but in the face of such huge losses (over 4 years) in the Maze, it would seem most improbable that increasing the numbers of cases would change the direction of a deficit of 30 T.Q. points.

Incidentally, two preoperative tests were given to both the experimental and control groups, and the practice effects were clearly apparent. Yet, notwithstanding the fact that the postoperative test was the *third* application of the same test, the losses in Maze scores were still quite marked. Again, as Hebb pointed out in discussing his own and others' results at the third research conference on psychosurgery in 1951, patients after psychosurgery are much more accessible to testing and are improved in affect. Increased cooperation should, he pointed out, increase the score. When, therefore, *deficits* increase after operation, the actual situation is probably

worse than it appears. The difficulty lies in the fact that there may be so many variables unevenly affected in a hospital population that an untreated group is no longer an adequate control. A simple example will suffice. In a study of the effects of chlorpromazine, the qualitative performance in the Maze was to be compared in an experimental and control group. But since hand tremors are either induced or increased by the drug, the qualitative scores of the treated group could no longer be used as points of comparison with the untreated. A factor extraneous to the purpose for which the test was applied had been introduced in the case of the experimental group and this hand tremor affected their scores adversely. Also, whenever placebo administration is involved, it is impossible to control suggestibility. To compare a hospital population with normals has its dangers, but patient-with-patient group comparisons are, in my opinion, much more uncertain.

As to when controls really control an experiment with humans has been thus briefly discussed to account for my apparent neglect of the comparisons made in what I regard as a methodology too often unwisely accepted as sound.

The third project in the Columbia-Greystone investigations, the New York Brain Study, was concerned mainly with the comparative effects of more anterior and posterior operations and, to a lesser degree, with the demonstration of practice effects. To this latter end, there were two preoperative tests and these showed for 44 subjects a practice gain of 1.79 years, or the equivalent of 12.7 T.Q. points. These figures should be compared with the Porteus-Peters practice gain for 55 prison inmates of 1.6 years or 11.4 T.Q. points. The first Columbia-Greystone control group gained 1.6 years, exactly the amount found in the Porteus-Peters study.

I have calculated from Sheer's [12] Table 4 the mean Maze loss of all 44 operated subjects to be 2.22 years or 15.86 T.Q. points, one of the largest deficits yet reported. But there were distinct differences when orbital and superior operatees were compared. The more superior operation resulted in a 20-point deficit ($N = 24$). Wechsler-Bellevue and Porteus Maze deficits are compared in Table 8.

Sheer sums up the findings of the two Columbia-Greystone projects and the New York Brain Study in relation to different degrees of ability to benefit from practice. Four subtests of the Wechsler-Bellevue Scale—Arithmetic, Similarities, Object Assembly, and Digit Symbol tests—showed in comparison with the controls a diminished ability to improve with practice in repeated test applications. To this conclusion he adds: "The two other tests which have also been consistently affected by the more posterior and

TABLE 8

T.Q. GAINS OR LOSSES AFTER PSYCHOSURGERY
NEW YORK BRAIN STUDY DATA

Operation	W.-B. Verbal	W.-B. Perf.	W.-B. Full	Porteus Maze
Orbital Topectomy	—1	—4.4	—1.0	—10.14
Superior Topectomy	—1.25	—7.37	—4.5	—20.6

superior operations are the Porteus Maze and Homograph tests." Moreover, their losses represented absolute deficits in spite of practice effects.

The Homograph Test presents a list of words, each capable of different meanings, such as file, note, club, band, and the subject responds with as many meanings as he can think of. Though the test is valuable in demonstrating loss in ability to "shift" after psychosurgery, it is so patently a test of verbal facility and general vocabulary range that we found it to be of poor discriminative value when applied to groups of patients in Hawaii whose English was restricted. As to the four Wechsler subtests, though the patients' ability to gain through repeated tests was diminished, their actual postoperative scores appeared small in comparison with Maze deficits.

CHAPTER REFERENCES

[1] Hutton, E. L., "Results of Prefrontal Leucotomy," *Lancet*, **i**, 362–66, 1943.
[2] Hunt, *loc. cit.*
[3] Rylander, Gosta, "Personality Changes after Operations on the Frontal Lobes," in *Acta psychiat. et neurol.* (London, Oxford Univ. Press, 1939).
[4] Partridge, Maurice, *Prefrontal leucotomy* (Springfield, Thomas, 1950).
[5] Ström-Olsen, R., Last, S. L., and Brody, M. B., "Results of prefrontal Leucotomy in 30 Cases of Mental Disorder," *J. Ment. Sci.*, 89, 165–81, 1943.
[6] Porteus, S. D., and Kepner, Richard, "Mental Changes after Bilateral Prefrontal Lobotomy," *Genet. Psychol. Monogs.* 29, 3–115, 1944.
[7] Porteus, S. D., and Peters, Henry, "Maze Test Validation and Psychosurgery," *Genet. Psychol. Monogs.*, 36, 3–86, 1947.
[8] *Proceedings of the Third Research Conference on Psychosurgery, U.S. Public Health Service* (Washington, Publ. 221).
[9] Smith, Aaron, "Changes in Porteus Maze Scores of Brain-Operated Schizophrenics after an Interval of Eight Years" (Paper given before East. Psychol. Assn., 1958).
[10] Petrie, Asenath, *Personality and the Frontal Lobes* (Philadelphia, Blakiston, 1952).
[11] King, Henry Eugene, *Selective Partial Ablation of the Frontal Cortex* (Hoeber, New York, Columbia-Greystone Associates, 1949), Chap. 14.
[12] Sheer, Daniel E., *Studies in Topectomy* (New York, Grune and Stratton, 1956), Chap. 6; *vide* p. 63.

CHAPTER VI

Further Maze Test Validation

Up to this point in the story of validating the Maze Test through psychosurgery, the question of the duration of the immediately observed postoperative deficits has been left open. Since more than one study has agreed in placing practice effects at 1.6 years of test age and the same investigations found a retest postoperative loss of 1.7 years for the second examination, it is apparent that the actual deficit is about 3.3 years. There is, as a matter of fact, considerable experimental support for this assumption; evidence also indicates that the loss tends to be permanent. Clinical experience has already been cited to show that the social effects of frontal lobe damage are considerable even though vaguely definable. Confirmation comes from the Columbia-Greystone projects.

Landis, Zubin, and Mettler [1] remark on the differences in personality after psychosurgery, "although sometimes the difference is not immediately recognized," thus bearing out the contention that both test and social consequences may be in part deferred. Speaking of the Greystone patients, they say: "They are apt to be somewhat more indolent; they are often outspoken, saying the first thing that comes into their heads rather than waiting to think what response the remark will produce in others. They are aware that they are hasty, undiplomatic, and tactless."

It is a great pity that such observations have not been systematized and based on long-term studies of frontal lobe patients, from Phineas Gage onward. All that we know of this classical "crowbar case" is what was reported by Dr. Harlow,[2] his attendant physician, who stated that while his employers formerly regarded him as a most efficient and capable foreman, he was unable after recovery to keep his job. Harlow stated:

He is fitful, irreverent, indulging at times in the grossest profanity (which was not previously his custom), manifesting but little deference to his fellows, impatient of restraint or advice when it conflicts with his

desires, at times pertinaciously obstinate yet capricious and vacillating, *devising many plans for future operation which are no sooner arranged than they are abandoned in turn for others appearing more feasible.* [Italics mine.]

Here, surely, is a picture of a socially incompetent person, his social handicaps being ineffective foresight and lack of planning capacity. He had suffered a very drastic accidental lobotomy, the instrument being a bar 3 feet, 7 inches in length, 1¼ inches in diameter, blown completely through his frontal lobes, entering below his left eye and emerging near the vertex of the skull near the line of the coronal suture.

Had this report of the social deficits been carefully studied, no one, least of all myself, should have been surprised at the impairment in Maze performance that commonly follows such injuries to the frontal lobes as are occasioned by lobotomy or topectomy. Now that brain surgery for the relief of psychoses has been almost completely discontinued in favor of chemotherapy by the use of tranquilizers, the opportunity to carry out adequate psychological studies before and after operation has been lost. There remains available for investigation only such patients who were operated upon years ago.

However, such studies based on reexaminations have a special value in providing evidence on the disputed point of the permanency of any resultant deficits. The first investigation was carried out by Dr. Mary F. Robinson [3] with 68 patients who had suffered prefrontal lobotomy, the average length of time between operation and examination being over 3 years. Unfortunately, she had no preoperative test scores and for her comparisons had to rely on a control group of 12 cases who were "former state hospital patients of similar age and diagnosis who had improved without operation." Five of these were employed, seven were keeping house.

Though this comparison with unoperated patients is open to obvious objections, it is noteworthy that while the controls reached 13.9 years in the Maze, about equal to the average normal population, the operated group scored only 10.6 years. This deficiency of 3.3 years is very considerable, equivalent to 23 T.Q. points. Since these patients had made "fair to good recoveries" and were reporting routinely to a follow-up out-patient clinic, mental deterioration was probably not an easily observable feature in their condition. The mean Maze deficit of 3.3 years is exactly what we obtained by adding the preoperative loss to the estimated practice gains.

Robinson also found significant differences as regards "deliberation" tested in various ways. Malmo, however, repeating her tests

on several patients, could not, in this respect, confirm the Robinson findings. Preoccupation with the self and concern with the past and future are, according to Robinson's results, also markedly affected by the operation. In the 10 cases that suffered a radical (more posterior) cut, the deficit in the Maze amounted to 4.3 years, or about 30 T.Q. points.

Fortunately, the opportunity for follow-up work with Columbia-Greystone patients has not been overlooked, and under the general direction of Landis and Kinder such a study has been carried out by Smith [4] and recently reported by him. The study provides conclusive evidence for the permanency of Maze deficits, particularly with regard to superior frontal operations. The interval of elapsed time between operation and reexamination was over 8 years. The subjects were 27 topectomy cases and 23 controls who formed part of the group of operatees and controls of the New York Brain Research Project which terminated in 1950. Eight of the experimental group and six of the controls, examined by the Maze, had been on tranquilizing drugs.

The 17 patients who underwent the superior cortical ablations had a mean Maze deficit of 3.88 years when their present scores were compared with their preoperative levels 8 years before. This loss is equivalent to almost 28 T.Q. points.

Those who suffered the apparently less drastic orbital operation ($N = 10$) had a much smaller mean deficit amounting to only .2 of a year, thus supporting the view that the focal area for planfulness as mediated by the frontal lobes is in the superior portions of areas 8, 9, and 10, and their neuroanatomically related structures. For the 23 controls, the loss was only .63 years or 4.5 T.Q. points.

The most interesting feature of Dr. Smith's investigation was the consistency of Maze losses in the case of patients with superior frontal lobe ablations. No less than 16 out of the 17 patients had Maze scores below their original preoperative Maze performance. In other words, 94 per cent of the group declined in score, the average deficit being about 3.5 years, equivalent to almost 25 points of T.Q. score, or 2 points more than Robinson found in her follow-up study. The average test age of these patients 8 years after operation is only 10 years, less than that of the Australian aborigines examined by the writer in 1934. In the face of this finding it would be futile to contend that damage to the superior portions of the frontal lobe has not occasioned most significant and permanent mental losses.

Through the courtesy of Smith, I am able to quote from a summary of his paper delivered to the Eastern Psychological Association this year. The following extracts are of special interest:

In our subjects, the superior operation (excisions of upper areas of frontal lobe tissue) showed greater and progressively deteriorating losses in successive postoperative examinations (immediately following and eight years later) than the orbital operation (lower frontal lobe areas excised).

Age of the subject is an important factor related to effects. Older operates showed greater losses in the follow-up examinations than younger operates in both types of topectomy.

Owing to the complex nature and dynamic properties of the experimental variable, brain damage, psychological measures may frequently reflect changes between initial and residual effects. . . . Our findings of increasing degeneration following the operations may thus have evidence of external validity for the psychological measures, although this cannot be definitely stated at present in the absence of autopsies of our cases.

The unique nature of the Porteus Maze Test and its greater sensitivity in discriminating the operates with superior topectomies from all other operates and controls is shown in comparison with the 11 subtests of the Wechsler-Bellevue, the Capps Homograph and Weigl Sorting Test. It also discriminated older subjects with operations from nonoperates much more reliably than the remaining measures.

Quoting now from Smith's main paper, he says: Thus of the four operate groups, three (the younger superiors, older superiors and older orbitals) do not show a "transient" loss followed by recovery and reported as a "drop and rise" pattern, but instead, a gradient of increasing losses. Only a single operate group, the younger orbitals, show the reported "drop and rise" pattern.

The results are shown in Table 9.

TABLE 9

PORTEUS MAZE SCORES AFTER TOPECTOMY
EIGHT-YEAR FOLLOW-UP STUDY

					Smith's Data		
Operative Groups	N	Preop. 2	Postop. 1	Loss (Yrs.)	Postop. 2	Loss (Yrs.)	T.Q. Loss
Superior	17	14.74	11.69	−3.05	10.0	−4.74	−34
Younger Ss.	12	15.54	12.59	−2.95	10.58	−4.96	−35
Older Ss.	5	12.80	9.70	−3.10	8.60	−4.20	−30
Orbital	10	11.55	8.28	−3.27	8.45	−3.10	−22
Younger Ss.	7	9.71	6.57	−3.14	7.57	−2.14	−15
Older Ss.	3	15.83	14.25	−1.58	10.50	−5.33	−38

Note: The above figures differ from Smith's table in that comparisons have been made between the second preoperative score, the first postoperative score, and the postoperative level eight years later. The large size of the first postoperative losses may be due to the fact that the subjects of study were the patients who remained in the hospital and represent the therapeutic failures.

Malmo and Penfield both reported greater losses for superior ablations. Moreover, von Monakow, Goldstein, and Meyer, as cited by Smith, have noted more marked and rapid degeneration in neurobiotactically linked superior regions, as contrasted with post-traumatic changes in the anterior portions of the frontal lobes.

As Smith rightly infers, these disparities of effect, especially those revealed by the Maze, deal a decisive blow to the theory of cortical equipotentiality. Landis [5] has summarized the behavioral effects of topectomy as loss in anxiety, zeal, and vigilance. I suggest that the anterior regions of the forebrain are more concerned with anxiety, while the superior areas mediate planfulness in which zeal or initiative and vigilance are main components.

The differences exhibited by younger and older orbital operates are in line with Meyer's observation of earlier and more rapid spread of degeneration in older cerebra after leucotomy. Smith suggests that if postmortem observations confirm these degenerative changes, these findings may provide external validation, both for the Maze and seven other measures which show significant though less consistent losses after an 8 year interval.

Landis made the first public mention of the above findings at a symposium devoted to an appraisal of the Porteus Maze at the 1957 A.P.A. meeting in New York.[6] Dr. Landis stated his opinion that the results of the Rockland Hospital follow-up study provided conclusive proof that the deficits that followed superior frontal lobe damage were permanent, a matter he had questioned in the 1951 Research Conference on Psychosurgery. Coming from such a high authority on the mental effects of psychosurgery, this acknowledgment should go far to settle the issue raised by the failure of tests to reflect mental changes after severe frontal lobe damages.

This evidence far outweighs the negative instances provided in a study by Ström-Olsen and his associates (11 subjects) and one by Jones,[7] in which the preoperative level of patients was 5 years, the "floor" of the tests. Obviously, the only change that could take place was in the direction of increased scores.

In an otherwise excellent summary of lobotomy results by Dr. Sidney Crown [8] of London, the author assigns, in my opinion, far too much importance to these findings, stating that the Ström-Olsen investigation "provides perhaps the main experimental support to the view that there are no significant changes in formal tests of intelligence following leucotomy." Crown, however, states that the conclusion "that the maze test is sensitive to postleucotomy changes in intelligence would appear justified." The accumulation of evidence with regard to psychosurgery would seem to indicate such caution in acceptance to be no longer necessary.

Though the validity of the Maze as a measure of practical social adjustment is no longer in doubt, some students may be misled by a statement that a test that is not reliable—in the sense that a score obtained today may not be the same tomorrow—cannot possibly be valid. This at once lands the student on the horns of a dilemma. Here is a test with unique validation as a frontal brain test that by test-retest correlation may have a low "coefficient of reliability."

It should be at once obvious that the relevancy of the finding depends on the nature and purpose of the test, and the method of obtaining the reliability index. The statement in relation to the Maze is about as meaningful as to say that a clinical thermometer is of no value because its reading varies with the same individual at different times. The variation is in the individual, not in the instrument. It is the method of obtaining the coefficient that is questionable. However, this is such an important distinction that it will be discussed more fully later.

Bearing in mind the principle that when the weight of evidence (due to the impossibility of securing a statistically sufficient number of cases) is far less than the experimenter would like to have, common sense does not scorn a partial answer. Several studies based on relatively few cases have such striking results that they cannot be dismissed as unimportant. The first refers to patients whose temporal lobes have been operated upon because of a special type of epileptic seizure. This investigation is being carried out and reported upon by a team of medical men and psychologists (Ian A. Brown, M.D., Lyle A. French, M.D., William S. Ogle, M.D., and psychologists Shirley Jahnson and Floyd W. Ayres). Patients were selected from a clinic conducted by the Division of Neurology at the University of Minnesota Hospitals.

As reported, there was some noticeable verbal impairment following temporal lobe surgery, but when tests were used "whose customary use is for the detection of intellectual deficit due to organic brain damage," no consistent deficits showed. This has a bearing on the theories of brain localization. Of the Porteus Maze the investigators say:

This is a test in which performance has been found to be adversely affected by frontal lobe surgery and which is thought to measure planning ability and foresight. All of four subjects tested before (operation), and 1 to 15 months after temporal lobectomy showed a decrease in Porteus Test Quotient (T.Q.), the average decrease being 24.25 T.Q. points. The significance of this loss may best be seen by comparing it to the average *gain* of 11 T.Q. points, which has been observed among normal subjects on retest after 4½ years, and to the smaller average *loss* of 12 T.Q. points, which has been observed among patients retested

within a matter of weeks after frontal lobe surgery. This Porteus T.Q. loss was observed in only 4 patients, but the magnitude and consistency of the loss recommends further work with this promising test.[9]

They further sum up the mental effects of the operation as follows:

The results of intelligence tests are consistent with the following conclusions. These patients do not come exclusively from any intellectual stratum; and their abilities vary from the mental defective to the very superior level of intelligence. Verbal impairment, which has been noted by other investigatores, characterized the majority of the patients tested in this series. Neither pre nor postoperatively did these patients consistently give test evidence of the kinds of intellectual deficit traditionally associated with organic brain damage. Porteus Maze Test results suggest that temporal lobectomy may result in impaired vigilance and foresight and that this impairment may be larger than the similar impairment which has been found to follow frontal lobe psychosurgery.

Though the above is by subtitle only a preliminary report, its findings lend support to my own suggestion [10] that maze-threading is mediated by both temporo-parietal and prefrontal regions. It also gives substance to the idea that verbal intelligence will be found to be represented posterior to the Sylvian fissure.

In the temporal lobe cases reported by Brenda Miller [11] in her review of this type of investigation she notes as one of the main results of her study of 25 of Penfield's cases an inferiority in "two visual tests (the McGill Picture Anomaly and the Wechsler Picture Arrangement), both of which involve complex picture material dealing with everyday social situations. . . . The second finding was the marked inferiority of the right temporal lobe group on tests of spatial patterning both visual and tactual."

Among earlier work, Miller refers to the Weisenberg-McBride psychometric study of 22 cases of lesions of the right side, which found that while these patients approach normality on some verbal tests, "they are significantly inferior on the arithmetic and most of the nonlanguage tests, notably the Porteus Maze Test."

This work makes plausible my suggestion that more than one region of the brain is involved in Maze Test performance and that parietal and temporal areas on the right side may mediate visual perception of simple space relationships that is needed for threading the simpler mazes. When the mazes become so complex that they call for higher level planning, the superior portions of the frontal lobes may come into play. That persons of higher mental levels tend to have greater losses, not only in the Maze but in other

mental measures, points to functional interaction of various neuroanatomical brain areas.

It must be admitted that this is a tentative theory with little experimental evidence. It was first put forward on the basis of observation of lobotomy cases at work on the Maze. They very frequently deal quite easily with the designs below the 11 years test age, but at about this point in the examination responses seem to change in character. "The subject seems to relax all vigilance, with errors, often repetitive, following in quick succession. . . . From an unhurried, easily confident individual, the subject suddenly changes to an unforeseeing, stupid, almost compulsive sort of person." [12] The explanation suggested is that with the temporal and parietal lobes intact the simpler mazes present no special problem, and with some subjects the switch to the frontal center is not necessary. Ordinary subjects, however, whose frontal areas have been damaged, fail with the more complex designs. The fact that Jahnson in the Minnesota study reports such a serious impairment as is indicated by a 24-point deficit in Maze T.Q. would seem to indicate that the temporal lobe plays an essential role in maze-threading. It should be noted that the deficit is almost as serious as that found in Columbia-Greystone II thalamotomy operations. I have also commented on the frequent absence in lobotomy patients of any evidence of emotional frustration after this failure, which is accepted with complete equanimity.

Whether or not this general theory of bifocal or manifold mediation is accepted, it seems clear that very widely distributed cortical areas, frontal and extrafrontal, are involved in Maze performance. There is also now considerable proof that the blood chemistry of the brain affects initiative and vigilance, which underlie planfulness. The first evidence on this point was dramatically presented in a study reported by O'Shea and others as long ago as 1942.[13]

Acting on observations that the vitamin-B deficient person "seems confused and apparently less able to reason and to exhibit good judgment under normal conditions," the authors set up a psychometric study of four subjects under an experimental diet together with four controls who remained on an adequate vitamin-B diet. After varying periods of time up to 98 days on the experimental or vitamin-free diet, the clinical picture (anorexia, easy fatigue, muscular tenderness) of vitamin deficiency was apparent in the experimental group, and at this point thiamine and B-complex feeding began.

In order to demonstrate intellectual impairment or otherwise, a series of tests was given during the period of deficiency and at two

other points in vitamin therapy. The list of tests included the Porteus Maze, the Tapping test, the Henmon-Nelson and Otis tests of mental ability, the Thorndike-McCall Reading test, a test for prose memory, and the Healy-Bronner Picture Recall. The Maze was applied alternatively in the standard and inverted positions, and a point-score system was arranged, the average score of 68.6 being attained on the standard test and 69.3 in the inverted position for all subjects, not a significant difference.

The authors summarize results with the Maze as follows:

> When deficient, the subjects' ability to solve the mazes was inferior to that of the controls, but following therapy with either thiamine or the B complex, their performance was equal to or better than that of the controls. [All differences were reported as statistically significant.]
>
> Furthermore, a definite relationship existed between the degree of deficiency and the degree of impairment of maze performance. The performance of the controls on the other hand showed no statistically significant variation throughout observation.

O'Shea, *et al.*, seemed fully cognizant of the relationship of these findings to social competency. They say:

> The exact extent to which the findings on the maze may be applied to problems of everyday life is still not certain, but one may assume that when in daily life a person deficient in the B vitamins must exercise the mental functions that are required to solve a maze, namely, foresight and judgment, he will do so less adequately than the person properly supplied with these factors.

The unique nature of the Maze Test was again demonstrated by the negative results achieved with the other tests. "General intelligence," say the authors, "reasoning ability (reading), and speed of hand muscle coordination (tapping) show no measurable improvement after therapy with thiamine or with the B complex." Nor were performances in these tests adversely affected during the vitamin deficiency.

It is noteworthy that this study was reported prior to the publication of any results with the Maze when applied to lobotomy patients, yet the findings are in substantial agreement with those that were obtained in psychosurgical experience. It is true that Guetzow and Brizek [14] tried to check the O'Shea-Elsom findings and did not obtain evidence of similar Maze impairment. However, their method of obviating practice effects was quite inadequate, consisting of 90 degrees' rotation of the four highest Maze designs. They also gave 10 applications of the test alternating examinations with

the rotated designs, with 5 applications of the Maze in standard positions. Practice effects in such circumstances would be relatively extreme. Their research design was obviously unsuited for its purpose.

Apart from the vitamin-deficiency study previously reported, another contribution to the measurement of effects in the way of positive therapy should be cited. Peters and Jones [15] have investigated, by means of Mirror Tracing and the Maze, the relationship of these test scores to improvement of male schizophrenic patients after psychotherapy, mainly psychodrama. There were 10 subjects in the experimental, 11 in the control group.

In the Maze there was a "substantial increase in score" for the experimental group after 3½ months of psychotherapy, the difference before and after therapy being significant beyond the one per cent level of confidence. With regard to the qualitative scores in the Maze, there was a significant reduction in qualitative errors for the experimental group, the change being insignificant in the control group. Peters and Jones remark, incidentally, on patient cooperation. They state that "one of the strong points of the Porteus Mazes is that only a very small number of hospitalized, psychotic patients fail to cooperate." A more complete account of this investigation is given by Jones and Peters [16] in a later paper. With the addition of several cases, the posttherapeutic picture as regards Porteus quantitative and qualitative score changes was still positive, the differences being statistically reliable. Those patients who accepted the Mirror Tracing test showed a marked tendency to improve. The Gardner Ward Behavior Chart also showed a significant degree of improvement. Of eight Rorschach factors examined, the differences were not statistically significant. Considering the evidence of change in social adjustment, this lack of relationship to the Rorschach is again surprising. The Draw-A-Man test changes were very slight or negative.

In regard to the relation between Maze Test scores and social adaptability, reference should again be made to the lobotomy findings discussed in the previous chapter. It will be remembered that in the Porteus-Peters study a definite pattern of test results was apparently associated with the degree of social recovery. The group that benefited most by the operation, to the point in many cases of apparent recovery from psychotic symptoms and hospital discharge, regained postoperative levels steadily and consistently until it exceeded its preoperative scores, and in four applications almost reached the ceiling of the test. In contrast, the unimproved group increased its scores very slowly and at no time reached much

above its preoperative level. The moderately improved group pursued a midway course of apparent recovery in Maze performance.

How much of these changes in level can be ascribed to practice and how much to improved social capacity is impossible to say, since both factors may be involved. But it is interesting to note that the psychologists and psychiatrists who took part in Columbia-Greystone Project I noted a similar relationship.

To define quantitatively the pattern of performance found by Porteus and Peters to be associated with social improvement, a criterion of patient selection was adopted by the Columbia-Greystone Associates. Six patients had a loss of one year or more at the first postoperative testing, which was regained at the second such testing, followed by a further gain. Six patients were also selected by the psychiatrists "as showing the greatest social improvement. All of them are at home and working." On an eight-point scale of improvement, these six averaged 5.83 points as against a mean of 1.0 point for all other operates, and 0.67 points for the controls. Five of these six patients exhibited the "drop-and-regain" test pattern described above. This situation may be interpreted as superior ability to profit by experience shown by recovered patients. This same factor apparently operates both in the Maze and in social adjustment.

In a report dealing with a follow-up within a year, Landis reported that one patient who exhibited the test pattern, but was not rated by the psychiatrists as among the most improved socially, was out of hospital within a year. His scores in one preoperative and four postoperative examinations were 11, 9, 11.5, 15, and 13. The drop of 2 years between the fourth and fifth examinations was not commented upon, but undoubtedly it had some significance.

As may be gathered, I hesitate to believe that psychosurgery improved the patient's social adaptability in a general, positive way but rather that it did so by diminishing psychotic behavior. I am of the opinion that social adjustment is furthered by the removal through psychosurgery of psychological hindrances in the way of the acceptability of the patient by his fellows. This might be called a negative gain rather than any strengthening of the patient's personality. An extremely malodorous breath may be a great social handicap. I should certainly not expect its correction, followed by social improvement, to be reflected by better Maze scores, unless the removal of a discouraging or depressing condition adds to self-confidence.

One other finding was interesting as confirming the Porteus-Peters results. It was that a performance pattern of a certain type

was "strongly associated with social improvement." This pattern was described as a distinct impairment of the preoperative Maze level followed by gains in the second and third applications of the test reaching or exceeding the preoperative level. This was the favorable predictive pattern found in the Columbia-Greystone Project I study. Unfortunately, the great decline in psychosurgical operations makes it unlikely that this lead will ever be adequately explored. Our own limited experience with greatly improved patients tends to confirm the finding that the pattern of initial loss followed by successive gains is associated with that improvement. Conversely, unimproved patients rarely show completely recovered Maze performance. Aaron Smith's study offers the most conclusive evidence on this point. His further follow-up examination of psychosurgery patients discharged from hospital will be awaited with great interest.

Noteworthy because of its attempt to set up independent criteria of validity is a study by Small.[17] He pointed out that studies of borderline intelligence, juvenile delinquents, and lobotomized patients have established the usefulness of the Porteus Maze in helping to distinguish socially capable from incapable people, and suggested that "it measures a temperamental or nonintellective aspect of intelligence function not assessed by standard intelligence scales."

But he goes on to say, "Not one study cited by Porteus, however, has directly tested the proposition. . . . In no instance does a validating study directly compare the Porteus with an independent measure defining foresight or planfulness." Small overlooks the fact that the Social Rating Scale previously described did test the Maze as being in part a temperamental or nonintellective measure, and that one trait-complex in that scale was initiative and planning.

Small used three scales of "foresight," one to measure "commonsense, everyday planning, the second a self-rating scale of planfulness, the third a personality inventory considered to reflect planning." In addition, he applied two standard Thurstone intelligence tests and an index of neuroticism. The planning scales were correlated with the intelligence tests, with the expectation that the correlations would not be significant as between intellective and nonintellective measures. On the other hand, they should correlate significantly with the Maze, if the latter really tested planfulness. The effects of neuroticism and intelligence were partialed out to see whether there was any significant residual correlation remaining. If so, Small considered that as proof of measurement by the Maze of nonintellective foresight. This procedure more or less cuts across

my position that planfulness is partly intellective and partly temperamental.

Small's subjects were 162 war veterans hospitalized at the Bronx V.A. hospital "for somatic, non-psychiatric complaints." It should be emphasized that the "foresight scales" represented the patients' self-judgments which are not necessarily true. Furthermore, these scales contain many items that seem to have only a very remote relation to foresight.

For example, one scale (AB) seems to me to be related to temperament rather than to foresight. Such statements as "I am easily embarrassed," or "I am happy most of the time," or "I am certainly lacking in self-confidence" do not indicate in any degree the presence or lack of foresight. In other words, the criterion itself must establish its claims to validity, otherwise the test may be a far better measure of planning than the criterion. However, accepting Small's measures at their face value, the outcome of the study is very interesting.

Small's summary of findings reads as follows:

The following results were obtained:

1. None of the planning scales was significantly related to either of the intelligence tests.
2. The Porteus Maze Test was significantly related to both intelligence tests, to all three planning scales, and to the Cornell Index (of psychological adjustment).
3. The Porteus Maze Test was significantly related to all these planning tests when the effects of verbal and non-verbal I.Q. scores were partialed out.
4. Partialing out the Cornell Index did not apprecibly affect the relationship between the Porteus and Scale P.S. (of commonsense everyday planning), but it did considerably affect the Porteus-AB correlation, raising it from –.223 to –.562 (zero order).

These results were interpreted as confirming and supporting Porteus's claim that the Mazes measure a personality or non-intellective characteristic of intelligence not encompassed by standard intelligence tests and that this trait could be called foresight.

It was held that the Porteus assesses *both* intellective and non-intellective elements in human functioning . . . and in so doing, provides an estimate of peoples' ability to use their intelligence in exercising practical, common-sense judgment. The possibility that poor mental health could act to suppress the effective interaction of intellective and nonintellective elements in the Porteus was pointed out in the interpretation of the sharp rise in the correlation

between the Mazes and Scale AB in which the effects of a common mental health factor was accounted for.

Suggestions for further research centered on making the Porteus a finer, less gross instrument for assessing a nonintellective planning element in human behavior. Suggestions for modifications of the test, structural analysis, and correlation with other tests of nonintellective aspects of behavior were made.

Among Small's suggestions for improving the Maze is one intended to obviate its criticism as a "thin" test by adding more mazes. This has since been done by the provision of the Extension Series (1955), and a third additional form to be published in 1959. If the psychologist is willing to apply at least two of these forms, chance successes and errors will be diminished, and will probably be eliminated altogether by the use of the three series.

Milton Jensen, like Small, suggested adding another more complex adult Maze. This could be done with advantage in the new, or third, form, but I have always been afraid to add it to a scale used with children since it may test a specialized ability. I would favor the devising of new, independent approaches to the measurement of planning capacity. Peters' circular mazes represent a somewhat different approach, but the provision of other means of measuring planfulness remains as a challenge to psychologists.

As to the clinical interpretation of Maze scores, one very important question concerns their relationship to ordinary industrial efficiency. With regard to feebleminded levels, correlations have been already quoted in Chapter III. These results were confirmed by Dr. Henry Herd,[18] medical officer to the Manchester (England) special schools, who obtained manual ability ratings from teachers of 130 children. Those rated "A" in ability had a mean Maze test quotient of 77, "B" of 74, "C" of 65, "D" of 59, while those rated lowest averaged 46. He commented as follows: "One notable fact emerged of some value for diagnosis: viz, that no child with a Porteus ratio below 50 made anything but a very poor appearance in manual work." Speaking of the tests in general, he said: "They therefore may be expected to, and do, correlate highly with manual capacity; obviously the qualities elicited, which are as much qualities of temperament as of intelligence, are indicators of probable social value."

The relationship of Maze scores to the industrial aptitude of defectives is thus established, but what is the situation with normal subjects where the picture is not affected by very low mentality or feeblemindedness? In 1929, Earle and Milner[19] reported from England on the value of the tests as prognosticators of the effects of

industrial training. Their subjects were apprentice fitters, electricians, and smiths. Progress ratings given by instructors halfway through the training course did not give significant results, but at the end of the training period their achievement correlated highest with the Maze in comparison with the other five tests applied. However, the coefficient was only 61, indicating the importance of factors other than planning ability in success.

But in 1952 the findings of a much more extensive study of normal male adults were reported by Lt. Col. Milton Jensen [20] in mimeographed form. He was then in charge of the mental testing of candidates for basic training at the Lackland Air Force Base in Texas.

The qualifying examination for retention in training prior to June 1952, consisted of the 6½-hour application of two tests (A.F.Q.T. and A.C.B.), and this procedure eliminated such large numbers of men, particularly Negroes, that Jensen decided to test them individually, using the Porteus Maze. On this basis it appeared that failure in the group-qualifying test was "closely related to indifference to, or aversion for, academic pursuits, and to psychological testing procedures." Evidently, this aversion did not extend to the Maze, for the distribution of Maze scores of 635 basic air-training examinees was closely parallel to the normal population's. This was in strong contrast to the A.F.Q.T. (Air Force Qualifying Test) in which all these men had I.Q.s below 75.[21]

Of 623 retested men, 511 with Maze T.Q.s averaging 104 (range 121 to 46), were retained in training, regardless of their qualifying test failures. Jensen stated:

> Scores of 633 male airmen retested with the Porteus Maze Test of Intelligence during the five months covered by this report evidence mental ability vastly superior to that indicated by their scores (or failures) on the group psychological tests. These findings (Porteus I.Q.s) are in general agreement with the clinical judgments of the psychologists and psychological examiners who have had close and reasonably prolonged association with these airmen in the retesting situation.

He adds the warning that this does not imply acceptance of the test as an adequate measure of mental ability in general. Several other tests were also applied and their results, to a lesser extent, were also taken into account in retention of men for basic training.

Our main concern is, however, with the training success achieved by these men who would normally have been excluded from such training. We quote Jensen's conclusion that, "individually retested men retained in Service are nearly as satisfactory in basic training,

according to training personnel, as are the other men in the flights in which they train. The differences probably are not statistically significant."

Actually 83.5 per cent of the men who passed the A.F.Q.T. were rated satisfactory, as against 71.5 per cent of the retested men. One of Jensen's comments is of particular interest. "Occasionally we find an airman for whom a low Porteus I.Q. is very unrepresentative. We have yet to find an airman who scored very high on the test (I.Q. 117 or higher) who did not provide other readily discernible evidence of intellectual power." He reaffirms this experience later by stating that

every individual scoring high on the Porteus Test presents qualitative evidence of power to plan ahead, to execute with precision and to adapt readily to tasks of increasing complexity. Once communication has been established, they emerge as vital, imaginative personalities, despite what we term cultural (including educational) deficiencies.

I am particularly interested in being able to cite Jensen's experience on this point, in contradistinction to my previously expressed opinion that failure in the Maze is more significant than success.

Jensen refers to those individuals with satisfactory Q scores as the "conformity minded" and gives a table of percentile scores for 532 trainees who failed the Air Force qualifying tests. The mean score for whites was 26 (S.D., 19), for Negroes 23 points (S.D., 19.8). These figures compare with our mean for High School sophomores and freshmen in Hawaii of 23 points, (S.D., 15.25). His percentile distribution agreed well with our own.

Chapter References

[1] Landis, Carney, Zubin, Joseph, and Mettler, F., "The Functions of the Human Frontal Lobes," *J. Psychol.*, 30, 123–38, 1950.

[2] Harlow, J. M., "Recovery from the Passage of an Iron Bar Through the Head," *Publ. Mass. Med. Soc.*, 2, 327, 1868.

[3] Robinson, Mary F., in Freeman and Watts, *Psychosurgery* (2d ed., 1950), Chap. XVII, p. 318 ff.

[4] Smith, *loc. cit.*

[5] Landis, *loc. cit.*, p. 5.

[6] "Symposium: An Appraisal of the Porteus Maze Test," 65th Convention A.P.A. (September 2, 1957).

[7] Jones, R. E., "Personality Changes in Psychotics Following Prefrontal Lobotomy," *J. Abn. Soc. Psychol.*, 44, 315–28, 1949.

[8] Crown, S., "Psychological Changes Following Prefrontal Lobotomy: A Review," *J. Ment. Sci.*, 97, 1–35, 1951.

[9] Brown, Ian A., French, Lyle A., Ogle, William S., and Jahnson, Shirley,

"Temporal Lobe Epilepsy; Its Clinical Manifestations and Surgical Treatment," *Medicine,* **35**, 425–59, 1956.

[10] Porteus, S. D., "Some Common-Sense Implications of Psychosurgery," *Brit. J. Med. Psychol.,* **28**, 167–76, 1955.

[11] Miller, Brenda, "Intellectual Functions of the Frontal Lobes," *Psych. Bull.,* **51**, 42–62, 1954.

[12] Porteus, S. D., *The Maze Test: Recent Advances* (Palo Alto, Calif., Pacific Books, 1955).

[13] O'Shea, Harriet E., Elsom, K. O'Shea, and Bigbe, Ruth V., "Studies of the B Vitamins in the Human Subject. IV; Mental changes in experimental deficiency," *Amer. J. Med. Sciences,* **203**, 388–97, 1942.

[14] Guetzow, H., and Brizek, J., "Intellectual Function and Reduced Intakes of B-Complex Vitamins," *Amer. J. Psychol.,* **59**, 358–81, 1946.

[15] Peters, Henry N., and Jones, Francis D., "Evaluation of Group Psychotherapy by Means of Performance Tests." *J. Consult. Psychol.,* **15**, 363–367, 1951.

[16] Jones, Francis D., and Peters, Henry N., "An Experimental Evaluation of Group Psychotherapy," *J. Abn. and Soc. Psychol.,* **47**, 345–53, 1952.

[17] Small, Kalman, "Planning as a Non-Intellective Component of Intelligent Behavior" (Unpublished thesis, Columbia University, N.Y., 1954).

[18] Herd, Henry, *The Diagnosis of Mental Deficiency* (London, Hodder and Stoughton, 1930).

[19] Earle, F. M., Milner, M., *et al., The Use of Performance Tests of Intelligence in Vocational Guidance* (London, H.M. Stationery Office, 1929).

[20] Jensen, M. B., "Basic Airmen Who Fail to Meet Minimum Mental Standards for Retention in the Air Force" (November 1952).

[21] Jensen, M. B., "Our Undeveloped Mental Resources" (Paper presented before Section I, Amer. Assn. Adv. Science, St. Louis, December 29, 1952).

CHAPTER VII

Qualitative Scores *

The foregoing brief review reveals a wide variety of Maze Test applications, all of which are reported quantitatively either in test ages or test quotients (T.Q.s). I use the latter term because I am convinced that the term I.Q., or intelligence quotient, when based on any existing scale of tests, implies an unjustifiable, therefore misleading, width of reference. If an I.Q. is quoted, it is commonly taken to mean a Binet or Wechsler-Bellevue age-test ratio. This implies first that intelligence has been properly defined and second that it has been adequately measured by applying such scales, an assumption that is far from completely true. It is demonstrably false as regards many of the feebleminded, the brain-damaged, people with inferior educational opportunities, people of primitive races, the nonscholastically minded, the psychotic, the criminal, or the delinquent. In such cases, life adjustment and educational adaptation cannot be equated, and to affirm that an I.Q. derived in the ordinary way is a sufficient index of social adaptability would fly squarely in the face of the facts; and if we are to define intelligence by what it does, we would be hard-pressed to conceive of it except in terms of mental adaptations underlying social adequacy. I am well aware that this would deny high intelligence to a man like Hitler, but under a definition that implies response to a range of stimuli of biological and social relevancy and in long-range planning, Hitler would not qualify. Only in the scholastic field is there anything like a close relation between success and a Binet or a Wechsler-Bellevue verbal T.Q., a relationship that may be expressed by a correlation of about .6. This means that the sum of the factors not measured by the test—and these include interest, ambition, ability to foresee and attain a goal, resistance to mental fatigue, sane reactions, and the other personal

* Portions of this chapter have been taken from *The Maze Test and Intelligence,* 1950.

characteristics we subsume under the term "drive"—are together almost as important as the traits the test measures. Possibly because there is this scholastic correlation, it is the schoolmen who bandy the term I.Q. about in staff meetings, attach unjustifiable importance to it in individual cases, and apparently regard with undue respect the rather mysterious processes by which it is obtained. Apparently social workers, who of all people have the best opportunities for observing what an I.Q. does not measure, have also fallen under the spell, judging by their readiness to accept a T.Q. definition of feeblemindedness. I am well aware that many people, educators and otherwise, would disclaim any such faith in the T.Q., but this is merely token opposition. Mental examiners are also quick to say that test results are always interpreted according to social history, but they never explain how this is done. If we refer to their reports, we see that their diagnoses are most often pegged to T.Q. standards. Psychologists would render an immediate service if they ceased once for all to speak of an intelligence quotient—which a test score is not—and substituted a test quotient, or T.Q., with the use of which no one can quarrel.

My view that the Maze Test measures planning capacity at a rather simple level, which is essential to every-day adjustment, would call for modification if clinical psychologists were limited to obtaining two scores only from the Maze: a test age and a derived test quotient. In the early years of the test's history the mistake of too-general reference was repeated. It was claimed that the Maze measured social adaptability, but this was plainly untenable in the face of occasionally excellent delinquent performance. The qualitative scoring, introduced in 1942, which was more concerned with temperamental defects than with intellectual incapacities, constituted a long step forward toward establishing the Maze as at least a partial index of social adaptation.

Extending the value of the test was, as a matter of fact, long overdue. If the temperamental factors affecting execution in the problem of maze-threading could be taken into account, they would go far toward changing the character of the test from one of planning to one of *planfulness.* In observing subjects at work, it was obvious that in many cases their ability to plan might be present but what they lacked was the prudent exercise of that ability. They were not functioning at their full potential capacity. This was particularly evident in the performance of many delinquents.

Attention was called very early in the test's history to the fact that two individuals may earn the same test score but differ widely

in the style or the quality of their performance. As far back as 1923 I wrote:

> It is noteworthy, however, that even though the test age credited to many delinquents may be relatively high, their mental alertness as shown in quickness of perception and readiness of action being often in evidence, their responses to the tests may show qualitative differences from those of normals of equal test age. In other words, observation of the test response may bring to light important temperamental differences which are not to be expressed in terms of mental or test age.[1]

More or less isolated studies such as those of Poull, Montgomery, and Karpeles, which already have been cited, showed that delinquents tended to have mean scores lower than those of well-adjusted children of similar social grade, but nevertheless many such subjects scored at or above average levels.

As another example, Dr. R. F. Jarrett[2] in 1926 reported in *The Lancet* a study of 100 youths in the English Borstal System committed mainly for theft or burglary. Their average Maze level was about 11 years 8 months, or about two years below the normal fourteen-year mean established by Cyril Burt in his English standardization of the test. Actually, over 10 per cent of these Borstal cases (aged 18 to 22 years) scored above 14 years, indicating a large overlap between delinquent and nondelinquent performance.

Purely as a matter of historical interest, it might be mentioned that as early as 1915 I examined a group of 22 delinquent boys in Australia and reported their average test age as being 2 years 4 months below their chronological age. At that time the test scoring was more difficult at the upper levels than it is at present, but regardless of the fact that 13 years was then the ceiling of the test, 25 per cent of the group scored "at age." At the time I commented on the impulsiveness of some of these subjects, and said that despite evidence of planning, execution was often faulty.

Nothing was done about scoring such errors for almost 25 years (except that in 1941[3] I contrasted the reactions of impulsive or poorly inhibited subjects with those of the mentally confused, irresolute type). In 1942, Dr. Charles Honzik and I began a special testing program with normal and "undependable" children, which later led to my working out[4] of a system that scored these qualitative defects in performance.

In the publication describing this work, the temperamental defects that led to poor execution were again stressed as against the intellectual weaknesses reflected in low test age. In order to discover these defects the subject should not know that his perform-

ance is being rated on any basis other than that of finding his way successfully through the Maze design. Thus his attention being mainly focused on the avoidance of blind alleys, habits of slovenly action, impulsive or haphazard performance show themselves more or less subconsciously. Conversely, the careful, dependable, conscientious person will likely display his characteristics by the way he works. These aspects of his personality were considered likely to conform with his ordinary style of response. This assumption of self-consistency and the setting up of personal standards of accomplishment have now been justified, as will be shown by more recent studies reported elsewhere in this volume. One other trait or trait complex, a compulsive perfectionism, was unfortunately overlooked in the 1942 scoring, as was also the tendency to make repetitive errors, a rather unusual but extremely important occurrence.

The position taken here is that personality is essentially "style of reaction" and that any job that we undertake is likely to reveal significant facets of character if they are accurately observed. The same assumption underlies the theory of projective techniques. The statement is particularly true if the task is not affected by some specialized skill. Drawing with a pencil through a series of Maze designs is certainly not affected by technical skill of any complexity. All subjects except those who suffer from some severe physical handicap operate on even terms in executing the Maze. The problem that confronted Dr. Honzik and me was discovering whether delinquents in a group show qualitative differences as compared with nondelinquents. The errors chosen for tabulation were cutting corners or crossing lines; lifting the pencil from the paper after a specific warning that this should not be done; taking a wrong direction but correcting it before a quantitative error would normally be scored; a general untidiness or waviness of line in the whole drawing. In regard to the last-named fault, samples of performance in each test have been published so that comparison of the quality of the drawing with standard performances could be made. These samples are to be found in the Appendix. Tracings equal to or worse than the examples are penalized as "wavy lines."

Because impulsive persons frequently make the first mistake possible, while, conversely, over-confident ones fall into error when they think the task is about completed, mistakes in the first third and the last third of the designs were also penalized. An additional weighting was given to qualitative errors made in the five- and

six-year designs, with which normally the testing begins. In the Extension series this feature applies to the seven-year test only.

The next step was to take a series of delinquent records from clinic files and tabulate the frequency with which the qualitative errors appeared. On this basis a weighting was assigned to each error.

The first trial of this weighted score, hereafter called the qualitative or Q-score, was carried out by tabulating the records of 100 delinquent boys, 100 delinquent girls, and 100 criminals. As a control group, Dr. Honzik and I examined normal high school students, 100 of each sex, and in addition 100 bus drivers employed by the Honolulu Rapid Transit Company. We first gave attention to discovering the frequency of these errors before any weighting of scores was allotted. We found that the total incidence of errors among delinquents was twice what was found in the performance of nondelinquent boys and girls. The criminals made more than three times as many errors as the bus drivers. There seemed to be no sex differences in Q-scores except that the girls demonstrated a greater tendency to disregard the instruction regarding lifting the pencil. This too was the most typical delinquent error. As a whole, delinquents are less attentive to rules and regulations and it was no surprise to find this tendency clearly reflected in the test performance.

Girls, whether delinquent or not, tended to be more careful and neat in execution which helped to counterbalance their greater tendency to disregard specific instructions. Our experience may be summed up by saying that in each and every qualitative error delinquents had a worse record than normals, with the gap in performance widening significantly with age.

When the errors were weighted and summated as Q-scores, delinquents' records were twice as bad as those of nondelinquent boys and girls of similar ages, while the average of criminal adults was three and a half times that of the bus drivers. Moreover the differences in Q-scores were entirely significant. For girls the critical ratio was 9.93, for boys 9.75, and for adults 13.3.

The question at once arose as to what influence the quantitative mental age scoring had on the Q-score. It could easily be that a good quantitative performance would be closely matched by a lower Q-score. This, however, was found not to be the case. For a group of delinquents the correlation between Q-score and test age was $-.40$ ($p.e. = .057$). For a group of criminals r equals $-.37$ ($p.e. = .058$). The Binet test age correlated $-.35$ ($p.e. = .059$)

with the quantitative score. Since a low Q-score indicated a good performance, a minus correlation represented the degree of agreement with mental age.

In order to make the Q-score usable in predicting delinquent behavior, I tried to establish a critical score above which 80 per cent of delinquents would be found. We found it to be 29 for boys; 32 for girls. In examining the criminals' records, we found that 86 per cent of them had Q-scores above this critical point. It was suggested that the critical score may be useful in differentiating the group of cases who become delinquent mainly because of environmental stresses and strains and may be, therefore, better prospects for reformative measures.

Two independent investigations seem to confirm our results. The fact that such a potentially useful measure had up to 1949 occasioned only two such studies could be in part attributed to the fact that war-time activities occupied the attention of a great number of clinical psychologists and specific research had to be deferred.

In 1944 Wright [5] scored the Maze qualitatively, using the test records of a group of 54 delinquent boys who were inmates of the Whittier State School in California. The mean Q-score of Wright's cases agreed almost exactly with that of 100 delinquent boys examined in Hawaii. I found that average to be 49.35 error points, while Wright obtained a figure of 49.26. The average Stanford-Binet T.Q. of her cases was 97.42 and the correlation between Binet and Q-scores —.28. It is evident from these results that the Binet T.Q. did not reflect the social maladjustment; on the other hand, considering that the nondelinquent Q-score is only 22 points, it is clear that the Q-score does differentiate delinquent performance. Moreover, the small r of —.28 indicates that the Q-score is only slightly affected by general intelligence as measured by the Binet.

The second study was by Grajales [6] and was carried out with adolescent delinquent boys, ranging in age from 13 to 16 years, who had been sent to Bellevue Hospital, New York, for psychiatric treatment. Their offenses included truancy, stealing, vagrancy, arson, fighting, sex misconduct, and so on, and nearly all were recidivists representing serious cases of delinquency. Seven of the group were defective mentally and an equal number were diagnosed as psychopathic. Grajales excluded the mental defectives from the first group of results.

Sixty of the New York boys were over 14 years old and therefore comparable with the cases studied by me in Hawaii. However, Grajales did not have any control group of nondelinquents and had

to rely for comparative purposes on the nondelinquent averages found in Hawaii.

As regards the delinquent cases, the Hawaii group had an average Q-score of 49 (S.D. = 25), while the New York group scored 56 error points (S.D. = 29). The difference between the New York and the Hawaii delinquents was not statistically reliable, but the difference between the delinquents' and the nondelinquents' performances of 22 points is significant as regards both the Hawaii and the New York averages. For this delinquent and nondelinquent difference the critical ratio is over 8.

Grajales also presents an analysis of individual errors and finds, as I did, that the tendency to lift the pencil in spite of the instructions not to do so, was the most frequent delinquent error. He suggests that his cases are more seriously delinquent than the Hawaiian group and that this may account for their worse average Q-score.

He also brings to light another interesting comparison. When he included defectives and 13-year-old boys to enlarge his group to 80 cases, the number of colored subjects increased to 25. His white delinquents averaged 61 Q-score points but the colored boys' mean was only 49, or exactly the average found in Hawaii. The majority of the delinquent group I examined were part-Hawaiian. It is also interesting to note that only 6 per cent of the New York delinquents had Q-scores better than the Hawaii nondelinquent average. The fact that 94 per cent had records that were worse than the nondelinquent average indicates the test's usefulness. It will be remembered that the nondelinquent weighted error score is 22, and Grajales remarks that only 5 of his cases had Q-scores below that average.

Incidentally, the correlation between the Wechsler-Bellevue and the quantitative Maze was .43. Grajales also quotes a correlation of —.22 between the quantitative and the qualitative Maze scores. It is quite evident from these figures that the two scores are not measuring the same complex of traits. That intelligence as measured by the Wechsler-Bellevue does not significantly affect Q-scores is shown by a correlation between the two of —.18.

Grajales summarizes his main findings as follows:

> Results of our investigation confirm Porteus' contention that his Maze used as a qualitative scale discriminates clearly between normal and delinquent adolescent boys. . . . The difference between the averages of these two groups is absolutely significant since the critical ratio is

8.6. . . . The average number of errors of our subjects was somewhat higher than that of Porteus' delinquent adolescents. This can be explained by the fact that most of our subjects were recidivists whereas Porteus' subjects were not.

With regard to the acceptability of the test, Grajales says: "The boys liked the test; they wanted to be tested. If at times we had to interrupt our work because of the requirements of the Hospital, or because of a visit of the parents, usually the boys came back willingly and without being called. Boys who have recently arrived at the Hospital asked to be given the test." That statement seems to dispose of the question of adequate motivation of his subjects.

Since I published the first study of Q-scores in 1942,[7] I have carried out another series of confirmatory tests. My object was to test the validity of the previous results by repeating experiments with other samples of both the delinquent and the nondelinquent adolescent population. We added some adult groups so that results on 20 additional groups were reported.

To check the previous average Q-scores, we retested 100 high school nondelinquents, and their average weighted error score fell from 22 to 19 points. In the earlier study our cases were drawn mainly from summer school students who were making up deficiencies in their work for the year. On the other hand the subjects of the 1942 investigation were regular session pupils.

We also examined an equal number of delinquent boys and girls. The 1941 weighted error score was 51.2 for both sexes combined but the 1942 cases averaged 55.4. Thus the nondelinquent average dropped and the delinquent score increased, widening the gap between nondelinquent and delinquent average performances from 28 to 36 points, thus increasing the significance of the difference.

To meet the objection that it is hardly fair to compare delinquents with high school students, we sampled 50 intermediate school children over 14 years old, who lived in a locality with a very high delinquency rate. Their Q-score average was 23 points, still two and a half times less than the delinquent mean, the difference being statistically significant.

We then asked teachers to select a group of students whose behavior in school was unsatisfactory, particularly from the standpoints of indifferent effort and undependable work in completing assignments. Twenty-eight such subjects scored 48 Q-score points, or only 7 points below the delinquent average. A similar experiment was carried out in the high school from which our control

group was drawn. Twenty-five of these students had unsatisfactory behavior records and scored 32 points, or nearly 13 points above the average of their more dependable fellow students.

These results were so striking that we repeated the experiment in another urban area where high delinquency rates prevailed. Fifty of the normally behaved students scored 22.5 weighted error points whereas 25 undependable students scored 48 points. The almost exact correspondence between these two sets of figures for two different areas of the city should be noted. All of these differences between normal and delinquent averages were statistically reliable. A considerable number of the unsatisfactory students were actually pre-delinquent, a fairly large percentage of them being later referred to the Clinic for conduct disorders. Thus at every point complete confirmation of the earlier results was obtained.

We then determined to check still further by examining another group of adult criminals. One hundred of these averaged 58 points, whereas the 1941 mean for a similar number was 56.7 points. It was difficult to find a group of comparative adults who might be used as a control group, and the only comparative results we can quote were the average scores of the bus drivers examined in the 1941 investigation. It will be remembered that a group of 100 scored about 18 weighted error points, or one-third of the score made by the criminals. Because some of these latter were Filipinos, we transferred our investigation to a sugar plantation. The Filipino workers there were mainly illiterate rural workers handicapped by language difficulties, whereas the prison Filipinos were city dwellers and English-speaking. In spite of their greater ability to understand instructions, the prison Filipinos averaged 60 points, the nondelinquent illiterate group 46.

In order to discover whether the scores reflected differences between satisfactory and unsatisfactory workers in industry, we examined two groups at a local pineapple cannery. Twenty-five male workers with satisfactory work records scored 14.3 points, the lowest mean weighted error score yet obtained. An equal number of workers who were reported by their supervisors as lazy, undependable, or slow in performance scored 37.3 qualitative error points. A group of women workers were also differentiated by their supervisors as dependable or the reverse, the more efficient group averaging 21.5 points, the less satisfactory 41.5 points, the difference again being statistically reliable.

Another possible source of error was also investigated. Because of our local conditions, Hawaiians and part-Hawaiians are overrep-

resented in both delinquent and near-delinquent groups. There might therefore be a possibility, though this is contradicted by the California and New York results, that what the Q-score indicated was not delinquent tendencies but the easygoing characteristics of part-Hawaiians. It has been observed that many of these people take very easily what we consider serious problems of life and go through no more pains than are absolutely necessary to support existence. That this attitude has a relation to delinquency is a popular assumption for which there is some evidence. Part-Hawaiian boys are often prone to exhibit little respect for others' property rights while girls slip easily into loose sexual relationships. Part-Hawaiian parents also often take their responsibilities in the matter of child control and training very lightly. Delinquency, therefore, is often not too seriously regarded by members of this racial group.

In Honolulu there is an extremely well-endowed school for children of Hawaiian ancestry which was set up in the days of the kingdom with funds left by Princess Bernice Pauahi Bishop. The properties left under this bequest have increased tremendously in value so that Kamehameha Schools have now excellent financial resources. The children who attend are, however, a selected group; a good deal of emphasis is placed on good character as well as satisfactory intelligence in the system of admissions. Excellent training is given, so that Kamehameha pupils are among the occupants of many responsible Territorial offices or positions. The amount of Hawaiian blood differs, as it is only necessary to prove even distant Hawaiian descent to be eligible for admission to the School. During four years' residence at school every attempt is made to correct careless or undependable behavior tendencies in pupils.

Whether it be due to rigid selection or the effect of environmental training, these part-Hawaiian pupils made an excellent record in the Maze Q-score. Thirty-one boys had a mean of 13.66 weighted error points while 25 girls averaged 15.8 points, the lowest for groups of both sexes yet reported. Comparison should be made with scores of 48.4 for delinquent boys and 62.3 for delinquent girls, the majority of whom were also part-Hawaiian.

As far as racial group averages are concerned, 219 nondelinquent Japanese students were first with 21.4 points; 72 Chinese scored 22.3 points; 83 part-Hawaiians, excluding Kamehameha pupils, were third with 29.5 points; 32 Portuguese came fourth with 34.6 points, and 24 Koreans were last with 35.4 points in Q-scores. Some

of these groups are obviously too small for proper group comparison and the differences in Q-scores need further confirmation.

In the previous 1941 study, critical scores of 29 for males and 32 for females were set up; the overlapping for delinquents was 20 per cent and 14 per cent for criminals. Again the 1942 study confirmed these results almost exactly, 86 per cent of criminals, 80 per cent of delinquent boys and 84 per cent delinquent girls being above the critical score. The overlapping of nondelinquent girls as regards critical scoring was 20 per cent; 21 per cent for nondelinquent boys.

The comparison of criminal groups classified by the nature of their offenses revealed some interesting differences in Q-scores. Those convicted of more violent crimes such as murder, rape, or assault with a dangerous weapon, averaged 67 points ($N = 36$). There were 59 convicted of less serious sex crimes, mainly intercourse with females under 16 years of age, and their average Q-score was 57.4 points. Burglars to the number of 55 scored 53.8 points, while 29 forgers and embezzlers had the best average with 44.3 points. Again the numbers in these various groups are rather small so that results of comparing them must be considered merely as suggestive; but as they stand they indicate a relation between the impulsive nature of the crime and the Q-score.

A rather confusing feature of the Q-score is the fact that it is based on weighted errors and therefore the higher the score the worse the performance, just the opposite of figures quoted for mental tests in general. To obviate this confusion, the scores of 826 nondelinquents examined in 1942 were arranged in a percentile table. In 1948, in order to check the accuracy of this table and incidentally the consistency of our results, 1,200 additional cases were scored and the results similarly arranged in percentiles.

TABLE 10

COMPARISON OF PERCENTILE SCORES OF TWO NONDELINQUENT GROUPS

N = 826			N = 1200		
1942 Score	1948 Score	Percentile	1942 Score	1948 Score	Percentile
90	102	0	16	15	60
46	47	10	13	12	70
35	36	20	10	9	80
28	29	30	7	6	90
24	24	40	1	0	100
20	19	50			

Table 10 gives the results of these two investigations side by side. The remarkably close correspondence between the two sets of percentile scores is an excellent demonstration of the reliability of the scores. It was noted in the preceding chapter that 532 basic airmen had an almost normal distribution of Q-scores according to our percentile table.

Probably one of the most careful confirmatory studies was reported by Docter and Winder [8] in 1954. They also examined a group of male delinquents ($N = 60$) at Whittier, California and compared their Q-scores with the performance of an equal number of nondelinquents "matched for age, sex, mental ability, race, and for the most part, socio-economic level." These were selected from the ninth and tenth grades of two California high schools and screened for delinquent tendencies by the school counselor and physical education teachers.

The mean Q-score of the delinquents was 47 as compared with 49 in the Hawaiian study, and for nondelinquents 25 as against our average of 23 points. The difference between delinquent and nondelinquent scores found by Docter and Winder was 22 points, with a critical ratio of more than 5.

They also found that 70 per cent of the delinquents examined scored above 29 points, as against 30 per cent of nondelinquents. This, they remark, is evidence that the Q-score "shows promise of predictive value." They further believe, on the basis of chi-square correspondence, that the weights assigned by me are reasonable.

However, they point out that the weighting is not essential, since the difference in unweighted errors has a critical ratio of 5.18, and a critical cut-off score of 16 yields about the same percentages above and below this score. Two scorers had a correlation of .98 between their markings, a very high reliability. The correlation of Q-scores with W.-B. was —.27. This, on the whole, confirms the writer's previous findings.

The latest study was reported by Fooks and Thomas [9] and was carried out with subjects in Connecticut institutions, matched with students from the New London High School on the basis of age, sex, intelligence, and socioeconomic level of parents. The delinquents had been diagnosed as psychopaths, or as having psychopathic tendencies, by psychiatrists. The authors compared scores by both the original Mazes and the practice-free Extension series, which is described in the next chapter.

On the whole, their results confirm those of previous studies, though they found a mean Q-score of 40 ($S.D. = 18$) for the delinquent group, the lowest yet recorded for that type of case.

The mean for their normals, however, was again 22 points (S.D. = 14.8) as found by me in 1942. The Extension mean Q-score was 46 (S.D. = 20) for the delinquents, 27.8 (S.D. = 17.3) for normals, the differences being significant beyond the .001 level of confidence. Errors therefore tended to increase in the Extension series for both groups.

Using the unweighted system recommended by Docter and Winder, with a critical score or cut-off of 16 Q-score points, the Connecticut findings showed that 74 per cent of the delinquents scored over 16 points in the original Maze, 84 per cent in the Extension. However, 28 per cent of the normals were above the cut-off in the standard Maze, 54 per cent in the Extension. Fooks and Thomas found no relation between quantitative or test age and Q-scores for either form of the test with delinquents, but a correlation of —.39 for the Extension with normals. They reported the same correlation (.98) between independent scorers as was found by Docter and Winder.

Two independent studies, interesting because of the possible relationship between neurotic and delinquent behavior, have been rather recently reported. The first of these was by Foulds [10] in England, who adopted Eysenck's [11] division of psychoneurotics into dysthymics on the one hand, and hysterics and psychopaths on the other. In terms of behavior, Eysenck describes the dysthymics as overscrupulous and inhibited; the hysterical, psychopathic individuals are underinhibited, shown by their tendency to transgress the rules of society by lying, stealing, sex delinquency, and avoidance of responsibility.

Foulds assumed that dysthymics should be distinguished by low Maze test scores—due not so much to impulsive behavior but to irresolution, psychomotor retardation, overconcern with errors—with generally good qualitative scores. In our own experience, this almost compulsive performance is often to be observed in Maze records of adult recidivists in prison. The psychopathic, on the other hand, are distinguished by heedless, slapdash performance, with disregard of instructions and a lighthearted attitude toward mistakes in the test. Very often there is considerable mental alertness indicated by average Maze test ages, but associated with extremely high qualitative error scores. Systematic observations of psychoneurotic traits, as revealed in Maze performance, should be most useful in interpreting scores that are often quite puzzling.

Following Eysenck's characterizations of psychoneurotic behavior, Foulds carried out a very illuminating study. He used a modified form of the qualitative or Q-score, taking into account

crossed lines, lifted pencils, wavy lines, and mistakes made in the first and last thirds of the test designs. His most important departure from standard procedures was to allow the subject to retrace his course after entering blind alleys and to score these errors under the category of "wrong directions." This is, of course, quite different to our own scoring of "wrong directions." This latter refers only to slight changes in direction of the course through the maze which do not quite involve actual entry into a blind alley but indicate an initial intention to do so. Fould's scoring is concerned only with actual entries.

Dysthymic symptoms in Maze performance seemed to Foulds to be vacillation, retardation, poor muscular control and coordination, inability to concentrate, and inferior recall of recently acquired experience. If these personality differences can thus be reflected in Maze reactions, then the test should be useful in differentiating the groups of psychoneurotics. Foulds theorized that affective disturbance implies preoccupation with self and that this awareness of one's own emotional state interferes with effectively carrying out a task such as the Maze provides. If, however, a quasi-automatic activity, such as imitative counting, is required of the dysthymic subject while he is threading the maze, his self-concern will be relegated to the background of consciousness and his performance will more closely resemble that of the hysteric or psychopath. Foulds also included in his data the time taken between giving instructions and beginning the test and for completing each design. His hypotheses were supported by the results of his study.

Briefly summarized, the records of 60 normals and 60 psychoneurotics showed that the former group took less time in starting and tracing and lifted their pencils less often. The latter group made more gross errors and had worse qualitative scores in the first two mazes—in other words, had inferior planning. If temperamental factors interfere in this way with the intellectual factor involved in success, "the test may well be expected to differentiate more closely between the psychoneurotic subgroups than between psychoneurotics and normals." (Generally speaking, correlations of the Maze with Raven's Progressive Matrices were noticeably higher for normals than psychoneurotics.)

Foulds' conclusion with regard to this part of the study was in line with the theoretical expectation: namely, that the test was found to differentiate more clearly between the psychoneurotic subgroups than between these subjects as a whole and normals. As examples, he cites for psychopathics prompt beginnings, rapid

execution with smooth flowing lines, impulsive errors, but less "lift pencils" than for dysthymics.

On the other hand, those with anxiety states start slowly, work fairly quickly, but with jerky or tremulous execution, and many "cross lines." Those with reactive depressions were slower throughout, made frequent errors in early stages and went further up blocked roads before realizing their mistake. (Scored as "gross errors.")

One finding of special interest in regard to our own work, to be reported later, referred to the performance of obsessionals. They are slow, score few gross errors, have fewer lifted pencils but make "firm, straight lines down the center of the channel and careful right-angled cornering with consequently very few crossed lines." We have found this kind of performance to be very noticeable with adult recidivists in prison but it is also observable among apparent normals. In other words, if it is characteristic of obsessionals, then compulsive behavior may lie in two directions, toward socially laudable and socially unacceptable reactions.

When Foulds retested psychoneurotic patients under ordinary conditions and then under rather mild distraction, he found the two groups clearly differentiated. The differences between the means were significant at the 5 per cent level of confidence for three scoring categories in the ordinary testing, but when distraction was introduced the differences were no longer significant. With regard to performance before and after E.C.T., Foulds concluded that the effect of distraction on Maze performance "is a useful method of assessing the intensity of an affective disturbance."

This study is described at some length, not only because of the actual findings but as an example of the possibility of fruitful studies in the field of abnormal behavior. My personal hope is that better methods of assessing the intensity of affective disturbance will be forthcoming through the provision of adequate rating scales. These will still be subjective but will be more susceptible to detailed analysis and reliability checks than are general diagnostic labels, such as dysthymic or psychopathic.

As regards the Q-score by the Maze, I agree with Foulds that the lumping together of scores related to dissimilar behavior traits may be open to objection, but it is, like a test quotient, a very convenient and practical over-all index of adjustability. His study points to the necessity for a more careful interpretation of items of performance that might constitute a far better scientific approach.

The second investigation to be cited also referred to neurotic

performance, but under conditions of stress rather than of distraction. It was carried out and reported by Aristotle Anthony [12] as a doctoral thesis in school psychology at Columbia University. In his introduction he speaks of the confirmation of differences in qualitative error scores between delinquents and others, particularly by Wright and by Docter and Winder, "who report a significant difference at the .001 level."

Anthony was interested in proving that increased tension raised the Q-scores of both normals and neurotics, but that the effects of stress would be apparent in a significantly higher Q-score for the neurotics.

His normal subjects were 30 teen-age boys attending the Ridgewood, N.J. public schools who had never been referred for either behavior or academic problems. The "neurotic" group ($N = 30$) had intelligence quotients above 90 but their school and over-all social behavior had required their being placed in an institution (the Devereux Schools) for therapeutic treatment. Both groups had been equated for age, intelligence, and socioeconomic level of parents.

The test procedure was first to administer the Maze under standard conditions. The second performance was under stress brought about first by allowing the subject to understand that his test performance had not been as good as it might be; second, by sounding a buzzer whenever he reached a choice point in the design; and third, by sounding the buzzer twice at one point which suggested to S that his work was so slow that he was in danger of receiving a zero score. Regardless of actual success, stress of this nature was imposed on all subjects.

The principal findings were:

1. Q-scores for both normals and neurotics were worse under stress. Beyond .001 level of confidence.
2. Q-scores of neurotics were worse than those of normals. Beyond .001 level of confidence.
3. Corresponding to Foulds' "wrong directions" score, more mazes (i.e. trials) were required under stress for both groups. Beyond .001 level of confidence.
4. Difference in Q-score between stress and non-stress was greater for neurotics than for normals. Beyond .05 level of confidence.

The Q-scoring applications were evidently very strict, as the mean Q-score for normals with no stress was 42 points (S.D. $= 24$), as against 78 (S.D. $= 32$) for the neurotics. Under stress the normals scored 55 (S.D. $= 25$), the neurotics 111 (S.D. $= 72$). It should be

remembered that the Q-score is an error score so that the higher it is the worse the record.

In Anthony's study, as was the case with Foulds' earlier investigation, the picture would be clearer if the terms neurotic or psychoneurotic could have been broken down into particular behavior traits, the degree of severity of which could then have been presented quantitatively. There are well-known objections to such behavior scales but in my opinion, their ratings are more meaningful than blanket psychiatric labels or general terms. As we well know, human judgments are fallible but that fallibility may be just as apparent in general as in particular descriptions of behavior.

Undoubtedly, a great deal of research is necessary before a weighted score derived from a summation of dissimilar errors in execution can be rightly interpreted. For example, the lifting of pencils may resemble crossing lines only in the fact that both represent a disregard of instructions, and this kind of heedlessness may be different from the impulsiveness shown by other qualitative errors at the beginning of a test design. The Q-score, therefore, is merely a convenient summary of the person's tendency toward inferior workmanship in the test. An investigation into the relationship of Q-score items to industrial efficiency would be very helpful and is the kind of study that, as indicated elsewhere, could best be carried out in institutions for feebleminded.

Chapter References

[1] Porteus, S. D., *Studies in Mental Deviations* (Vineland, N.J., Research Dept. Publications, 1922).

[2] Jarrett, R. F., "Some Observations on Social Capacity: Application of the Porteus Maze Tests to 100 Borstal Lads," *The Lancet*, **20** (November 1926).

[3] Porteus, S. D., *The Practice of Clinical Psychology* (New York, American Book Co., 1941).

[4] Porteus, S. D., *Qualitative Performance in the Maze Test* (Vineland, N.J., Smith Printing House, 1942).

[5] Wright, C., "The Qualitative Performance of Delinquent Boys on the Porteus Maze Test," *J. Consult. Psychol.*, **8**, 24–26, 1944.

[6] Grajales, Manuel L., "Porteus Qualitative Maze Test as a Measure of Delinquency" (Unpublished thesis, Fordham University, 1948).

[7] Porteus, S. D., "Q-Scores, Temperament and Delinquency," *J. Soc. Psychol.*, **21**, 81–103, 1945.

[8] Docter, R. F., and Winder, C. L., "Delinquent *vs.* Nondelinquent Performance on the Porteus Qualitative Maze Test," *J. Consult. Psychol.*, **18**, 71–73, 1954.

[9] Fooks, Gilbert, and Thomas, Ross R., "Differential Qualitative Performance of Delinquents on the Porteus Maze," *J. Consult. Psychol.*, **21**, 351–53, 1957.

[10] Foulds, G. A., "Temperamental Differences in Maze Performance," *Brit. J. Psychol.*, **42**, 43, 209–17; 33–41, 1951.

[11] Eysenck, A. J., "A Dynamic Theory of Anxiety and Hysteria," *J. Ment. Sci.*, **101**, 128–51, 1955.

[12] Anthony, Aristotle, "Normal and Neurotic Qualitative Porteus Maze Performance Under Stress and Non-Stress" (Unpublished Ph.D. thesis, 1958).

CHAPTER VIII

Extension Maze Series

During its long history, the Porteus Maze Test has come under considerable, but not unexpected, criticism. In following the commonsense practice of using the tools that are available, even though imperfect, psychologists have tended to overlook the shortcomings of their most familiar instruments. Demands for information from the use of such clinical standbys as the Binet and the Wechsler-Bellevue have been so insistent that rigorous criticism of these measures has been lacking. Even the Maze has gained in examiners' confidence through long usage. In other words, validity has been accepted on insufficient evidence. A factor in this situation is the time and effort necessary to set up external criteria for validation purposes. Then too, psychologists have realized that more prestige can be gained by devising new tests than by improving the old ones.

Another difficulty has been the truly tremendous multiplication of testing devices. Considering the paucity of workers devoted to research in this field, this matter of testing the tests gets scant attention. In any case, time attends to the vast majority of these new ventures in the psychometric field; unfed by continued research, they wither on the vine.

But this dearth of critical examination should not apply to the hardy survivors, of which the Porteus Maze is one of the most durable. For them, if a keen cutting edge is to be maintained, the fire of criticism is an essential stage of the tempering process. But the laborers have been few. Mazes are connected in many psychologists' minds with animal experimentation. The idea that much good can come out of pencil-and-paper labyrinths, as applied to human beings, seemed to many academic psychologists very doubtful. Others are so wedded to the use of the Binet or Wechsler-Bellevue that a serious disparity between these test results and the Maze is to them merely confusing. They have remarked: "We know that the Maze has had some extraordinary validation, but we don't know what its scores mean."

One of the most critical surveys of the Maze, or of any test, was carried out by Tizard [1] and appeared in a British journal. It prompted my reply that duly acknowledged the test's imperfections, mostly sins of omission. Where special pleading on behalf of a testing device is called for, I would have no hesitation in using it, but not to the extent of hiding its faults. What the critics overlook is the fact that these faults seem to be common to all tests. Shortcomings of the Maze are admitted, particularly in one direction, as set forth in the following extract from my reply [2] to Tizard:

> The most glaring deficiency is the lack of a supplementary series, which should be devised, standardized, and applied to delinquents, defectives, psychosurgical cases, and primitive ethnic groups; in short, in all areas where the original Maze has proved useful. Its provision, however, should not, as Tizard suggests, be for the dubious purpose of test-retest coefficients, but to improve the validity of a test series which is too brief and, therefore, affected occasionally by accidental success or failure. I would agree that the data available on *g*-factor loading, and, indeed, on factor analysis involving the Maze Test, reveal for those who like these purely mathematical approaches an "astonishing paucity of research." The differences between Maze test norms for peoples of different cultures are, as Tizard points out, 'unclear,' and someone should repeat my studies in the Kalahari, Central Australia, and South-East Asia, or wherever the peripatetic psychologist can lay hold of primitive peoples. It is true also that standard deviations for individual items of the qualitative score have not been provided, and no doubt the multiple correlation method would have improved their weighting. Likewise, it would undoubtedly be an improvement if percentile tables had been provided for each mental age in the test, as Tizard suggests. Moreover, consistency of scoring by different observers has not been properly checked, and I would agree that the Maze needs more systematic study both as an intelligence test and as a measure of social adaptation, though I should not separate those purposes. Certainly, much more research should be expended on such a popular test.

What seems to me to be the most obvious inadequacy in this list of defects has now been corrected, namely as regards a supplementary series, to be described in this chapter. But as is so often the case, this new approach has set up its own research problems, increasing rather than diminishing the list of things to be done. One last consideration explains Tizard's most serious criticism. Any person has only limited funds and time to devote to research, and certainly no one can say that the author of the Maze has spent too little effort in persuading his colleagues that the Maze Test is well worth attentive study. When the workers in the vineyard are

so few, the cultivation may well be shallow and the grapes sour. Each new application of the test among its great variety of uses seems to uncover an added flock of questions demanding clarification or solution. This confusion can only be dispelled by more research.

Obviously, if clinical psychologists could check the results of the original Maze against another approximately equivalent Extension series, their faith in the instrument should be greatly strengthened. No one is likely to doubt that two measurements of the same set of traits are better than one. If the two series in general yield the same scores, that fact would be useful in clinical appraisal of individual performance. For if it appears that an individual deviates markedly from the usual pattern of close agreement of the two scores, that fact may be of great clinical significance.

Clearly, individual deviations can be expected to appear quite often, since one purpose of the test is to measure the degree to which the subject profits by experience, obviously one of the ways in which humans differ very greatly. It must be kept in mind that the nature of the problem does not change throughout the series. Over and over again S must decide whether he shall turn right or left, enter or pass by an opening. Under the system of allotting repeated trials, there are over 100 possible choices in working throughout the whole set of mazes, though some of the traps are so obvious that scarcely any person with a minimum of maze-threading ability enters them. They add, however, to the over-all complexity of the design. The important point is that the Maze constitutes a general learning though not an item-memorizing experience—and individuals notoriously differ in learning ability. Some people never learn adequate caution, no matter what their mistakes cost them. There is a special need for research to assist in interpreting declines in score in the Extension series as compared with the standard or Vineland Revision.

The first methodological difficulty was in the actual devising of the Extension series, bearing in mind that it was not to be an alternative but a supplementary form. An entirely novel plan of having all the mazes circular in shape was suggested, but it was felt that there were advantages in not changing the subject's orientation to the task, already adopted through familiarity with the general design of the original mazes. Undoubtedly, the difficulty of circular mazes would be increased, since eye movements involved in scanning circular designs are different from those used in straight-line mazes. In the latter, corners are easily recognizable landmarks, making the identification and memorization of courses

after preview much easier. A former colleague, Dr. Peters, has experimented with circular maze designs but his scale needs standardization.

Then, too, a complete change in general design would require hundreds and probably thousands of cases and many adjustive changes before standardization could be achieved and its relation to scores on the original series determined. If the mazes were printed mirror-wise, the scale would be changed materially, and with the original tests as a basis, all that would be needed to achieve equivalence would be to lengthen certain alleyways in the maze or to add slightly to the number of traps. This suggestion, offered by J. R. Porteus, was followed. After trial application to a group of cases, it was found that a close approximation in group scores resulted.

It was recognized that the faults in standardization which belonged with the original series would be repeated in the extension. If there were not too much difficulty in obtaining equivalent scores, that would be indirect proof that graduated mazes could be regarded as consistent measures. Success in matching the scores of comparatively small groups would further indicate that variations in maze-threading skill, even if marked, tend to be spread evenly through the population in which the test is being standardized.

Instead of a huge single investigation requiring thousands of cases, I chose to give the study the character and force of repeated experimentation by breaking up the cases into smaller groups, then tabulating and comparing results. The constitution of these groups was based on order of examination of cases in each separate school, and to that degree represented random selection. In an attempt to discover how small in size groups could be in order to yield reasonably equivalent scores, the cases were compared in groups of 25, 50 and 100.*

Our results showed that approach to equivalence of scores in the two forms was achieved by comparing subjects of the same age, educational and social status, collected into groups of 50 cases, but that 25 cases gave scores only approximately similar. This fact may be important in regard to research involving applications of the two Maze forms to small groups. By dividing the cases I believed that reliability of findings would be just as assured as if any theoretical or statistical formula was applied. If 300 cases were available, considering them in successive fifties would surely

---

* Much of the succeeding material in this chapter has been reproduced from *The Porteus Maze Test: Recent advances,* 1955.

have the force of repeating the experiment six times.

It should be explained here that the study was undertaken for two specialized purposes: the first to throw light on the reactions of delinquents as compared with nondelinquents; the second to provide a relatively practice-free form of the Maze that would be available for retesting psychosurgical cases before and after operation, or in other research where control of practice effect was necessary.

In each case the tests were given on the same day, the Extension series being applied immediately after the original. Testing instructions were not repeated when the Extension was presented. The average time taken by each subject for the two series was about twenty-five minutes. No individual complained of fatigue and very few showed any reactions of relief when the testing was over. The examiners were probably much more fatigued than their subjects. Though the means for each group of twenty-five cases of the same sex were not relied upon for the purposes of comparison, they are given in Table 11, along with the means by successive fifties and hundreds.

TABLE 11

ORIGINAL MAZE AND EXTENSION—BY SUCCESSIVE GROUPS

		ORIGINAL			EXTENSION			DIFFERENCES		
School	Sex	N=25	N=50	N=100	N=25	N=50	N=100	N=25	N=50	N=100
McKinley	M	15.72			15.68			−.04		
	F	14.82	15.27		14.7	15.19		−.12	−.08	
	M	14.8			15.18			+.38		
	F	15.06	14.93	15.1	14.62	14.90	15.045	−.44	−.03	−.055
Washington	M	15.48			15.64			+.16		
	F	15.48	15.48		15.66	15.65		+.18	+.17	
Leilehua	M	15.9			16.0			+.1		
	F	15.34	15.62	15.55	15.0	15.48	15.575	−.34	−.14	+.025
Farrington	M	15.76			16.16			+.4		
	F	15.56	15.66		15.2	15.68		−.36	+.02	
Roosevelt	M	15.7			15.98			+.28		
	F	15.26	15.48	15.58	15.32	15.65	15.665	+.06	+.17	+.085
ALL N = 300				15.41			15.43			+.02

The agreement between the scores on the two series of tests is seen to be remarkably close. The largest deviation between groups of twenty-five subjects was .44 of a year of test age, the average difference in score for the twelve groups being .24 of a year or 2.9 months of mental age. The cases were grouped and tabulated in the order in which they were tested and to this extent represent random sampling.

When considered in groups of fifty, the largest difference between original and extended Mazes was .17 of a year, the average deviation being a little over .1 of a year, equivalent to 1.22 months of test age, while the disparity was reduced to .02 of a year for the whole group of 300 cases. Under these circumstances, it seemed quite unnecessary to figure critical ratios, regardless of the standard errors of the differences. The means and S.D.s for all cases in successive hundreds are given in Table 12.

TABLE 12

MEANS AND S.D.S—ORIGINAL AND EXTENSION MAZES

		ORIGINAL		EXTENSION	
N	School	Mean	S.D.	Mean	S.D.
100	McKinley (Washington)	15.1	1.755	15.045	1.72
100	(Leilehua) (Farrington)	15.55	1.49	15.575	1.46
100	(Roosevelt)	15.58	1.365	15.665	1.5
300	All Schools	15.41	1.506	15.43	1.6

The purpose of dividing the sexes equally was to obviate the effects of any sex differences. In one of the earliest studies of the Maze (old scoring) involving 1,000 cases, boys excelled girls at all ages from 6 years onward, with the exception of two groups at 12 and 13 years. In a second group of 1,255 cases, girls were superior at 12 years only. The data were published in 1920.[3]

Using only the 14-year and two adult tests, Porteus and Bassett [4] examined high school students in Vineland, New Jersey, and concluded that boys between 14 and 17 years tended to excel girls both in speed and accuracy. The comparisons, however, could not be reported in terms of test ages.

In a study of cases on Saipan by Joseph and Murray,[5] recently reported, the investigators found that boys excelled girls at each

age level in both Chamorro and Carolinian groups, the two ethnic strains represented in the island population. As regards our present study, 150 boys averaged 15.56 years on the Original Maze, as against 15.25 years for the same number of girls. In the Extension the means were 15.77 years for boys, 15.08 years for girls, a somewhat larger advantage in the case of males. Table 13 gives results by successive groups of 50 and the age level differences for each. The standard deviations for the groups as a whole ($N = 150$ for each sex) and the application of the ordinary formula gives a critical ratio for the original Maze differences of 1.8, just short of complete statistical significance. However, taking into account the fact that male superiority has been demonstrated in four previous investigations and holds good for both civilized and primitive peoples, it would seem to be very well authenticated. As regards sex differences in performance on the Extension, the critical ratio is 3.8, indicating complete statistical significance.

Turning now to the relation of the original and Extension Maze by separate groups (Columns 2, 5, 8, and 9 of Table 13), we find

TABLE 13

ORIGINAL AND EXTENSION MAZE MEANS—SEX DIFFERENCES

No.	Age M. Orig.	Age F. Orig.	Diff.	Age M. Ext.	Age F. Ext.	Diff.	M. Diff. Or. & Ext.	F. Diff. Or. & Ext.
50	15.26	14.94	M+.32	15.43	14.66	M+.77	+.17	−.28
50	15.69	15.41	M+.28	15.82	15.33	M+.49	+.13	−.08
50	15.73	15.41	M+.32	16.07	15.26	M+.81	+.34	−.15
150	15.56	15.25	M+.31	15.77	15.08	M+.69	+.21	−.17
150	S.D.1.415	S.D.1.58	CR.1.8	S.D.1.3	S.D.1.76	CR.3.8		

NOTE: Column 4 gives the differences between males and females in the original series; column 7 for the extension series; columns 8 and 9 compare original and extension scores by sexes.

that the largest difference was about one third of a year in favor of the boys in one group of 50, and that each group of boys did better on the Extension, the difference for the whole group ($N = 150$) being .21 of a year or three months of test age. With girls the differences between original and Extension mean scores for successive groups of 50 each were on the whole smaller, but all in the reverse direction, their tendency being to score lower on the Extension, the deficiency amounting to .17 of a year or 2 months of test age.

Though in this study the Maze Extension has been applied to subjects 14 years and over but not differentiated into age groups, it would be interesting to divide them by chronological age, not only to provide an additional check on the equivalence of scores, but also to help in determining the average age to be used as a divisor if age-test ratios are calculated; even though, on the basis of long experience, I question somewhat the exact value of the so-called T.Q. when applied at adult levels.

In the case of the Maze, the ceiling of the test is a score of 17 years if one adult test is given, or 19 years if two adult tests are applied, as in the two Columbia-Greystone studies. There is no difficulty in devising more complex mazes, but they can become so complex that only very extraordinary forethought and memory for spatial relations will avail in their solution. In short, the test takes on the nature of a puzzle that requires very specific abilities. One might as well use an end problem in chess to measure foresight involving spatial relations. The differentiation of ability of this kind and at this level would have little or no social significance except in a population of chess players.

Thus, setting a score ceiling at the comparatively low level of 17 years concentrates scores at the upper levels; and the older and brighter the subjects, the less normal the distribution of intelligence quotients.

In view of the above considerations, I present the data with regard to separate age levels above 14 years with considerable dubiety. Proper sampling for age standardization under these conditions presents extreme difficulties. I have already pointed out how this may affect results obtained with 14-year-olds examined in high schools. The same difficulty attends the representation of 17-year-olds in the present body of data. Since 11th- and 12th-year high school students are not proportionately represented, students in the 17-year group may be duller than average students. Yet how does the examiner obtain a representative sample of the population at above elementary school levels unless he has the power to go out into the highways and byways and seize for examination all the adolescents he meets? Obviously, our figures for high school students prove the equivalence of the original and the Extension series for that special segment of the population only. However, there is good evidence for assuming that the Maze is less affected by cultural level than any other test in common use, and cultural level includes education. With such reservations in mind, the scores by age groups are presented in Table 14. For the three largest groups the difference in test ages between the original and

TABLE 14

TEST AGES BY AGE GROUPS

Age	N	Original	S.D.	Extension	S.D.	Difference
14	88	15.72	1.31	15.74	1.39	+.02
15	80	15.26	1.69	15.18	1.76	−.08
16	70	15.23	1.49	15.14	1.69	−.09
17	39	15.15	1.8	15.43	1.63	+.28

Extension Maze series is again less than .1 of a year. Only in the relatively small group of unrepresentative 17-year students does the difference rise to .28 of a year or 3.4 months of test age. Thus, no matter how the data are grouped—by schools, by sexes, by ages—the near equivalence of the test scores can be demonstrated. If considered as an age scale, it will be seen that the 14-year-old's exceed the older children in both the original and Extension, but this superiority may very well be due to sampling conditions. Probably 14 years can be considered the general average of the population. In an earlier form of the tests, Burt's [6] English subjects, averaging 14.2 years, scored 13.8 years, while my own studies also showed a decline in average score beginning at 12 years, when studies were confined to elementary school children. The present study shows no consistent increase of score beyond 14 years.

Taking results as they stand, the success with which practical equivalence of score has been attained has been very surprising. Any very close approximation of scores on the basis of only 300 cases was certainly far from my expectation. I believe that the explanation lies in the way the new series was devised.

But a study of the records brought to light a most important though not unexpected result. Though the means were practically equivalent, individual performances in both series were in some cases far from being equal. This, of course, is of paramount significance from the clinical standpoint. If, for example, there were a perfect or even a very high correlation between scores in the two series, the new tests would have nothing but confirmatory value, since the original test would do all that the Extension could do also. What the psychologist needs to measure is the degree to which individuals improve or deteriorate with further testing. Fortunately, when a correlation was worked out for a group of 50 cases, the coefficient was only .5, high enough to indicate some general agreement, but low enough to show that individual deviations were fairly frequently apparent.

How large and how frequent these deviations are was revealed by the following tabulation:

1 year or less difference	67.6% of cases
1½ years or less difference	82 % of cases
2 years or less difference	89.3% of cases
More than 2 years difference	10.6% of cases

A difference of one year or less in the two scores cannot be considered of any great significance. It affects roughly two-thirds of all cases. Whether a difference of more than 1½ years affecting 18 per cent of the cases is very significant, only clinical research will prove. But undoubtedly the approximately 11 per cent of cases who differ more than 2 years in their scores demand close clinical scrutiny. When the records of these cases are still further analysed, it appears that 5 per cent improve their scores, leaving 5.6 per cent who lose ground in their second examination by the Maze.

Obviously, any responsible person interested in the adjustment of youth, be he educator, industrial trainer, probation officer, social worker, or juvenile court judge, would welcome from the psychologist a reliable estimate of an individual's stability or dependability, especially his inner compulsion to do a good job. Moreover, if the administrator is interested in the rehabilitation of wayward youth, he would be concerned with the detection of the "new broom" type, the individual who cannot stay the course and whose good intentions are evanescent. However, conformity, if expressed in purposeless or socially inacceptable fashion, may become obsessional.

Thus, personalities at the very poles of adaptive behavior have been differentiated more closely by the use of the two series of Mazes. This purpose is greatly furthered if both the quantitative and qualitative scores are considered. Watching subjects at work, the psychologist can observe one subject becoming more confused with errors and unable to cope with a lengthier or more difficult task; another who exaggerates its complexity and betrays a great lack of self-confidence; a third who becomes overconfident and dashes impulsively ahead or who relaxes vigilance just when the end is in sight and makes qualitative errors. Still another apologizes for his mistakes and proceeds at once to make more. Then there is the perfectionist who is so intent on extreme neatness and in avoiding crossed lines that he loses sight of the main purpose of finding his way out of the maze. The real question is, of course, whether these responses constitute a true sample of behavior, whether the attitude to the test is characteristic of the individual's usual reac-

tions. These points have already been discussed. Whatever doubts there may be on this score, I believe will be greatly lessened by using the Extension as well as the original Mazes. New evidence on this matter of representative performance is presented and discussed in a later chapter.

To anyone having experience in the social adjustments of youth, the similarity of the test responses with delinquent behavior described in the preceding paragraphs will at once be apparent. They are obviously the handicaps that dog the steps of wayward children. Many of these "relinquents" are long on promises, short on performance. They seem eager to keep out of trouble but are either unable to keep up the necessary effort or are too forgetful of past experiences to avoid repeated delinquency. Some are like runners who do not husband their strength and cannot put out that extra effort in the last stretch, which makes the difference between victory and defeat. The most that we can claim for the Maze is that it seems to come nearer reflecting such characteristics than any other available test.

Because of the extreme range of possible scores, from 0 to 100, which are possible in the Q-scores, and the fact that they are drawn from performance on different items, wide variability can be expected. Since there is not a high correlation between test age and Q-score, the largest coefficient reported being .44, the near-equivalence that has been demonstrated for test ages will not argue a like condition for qualitative scores. It should be remembered that in arranging the Extension a deliberate attempt was made to equate the two sets of mental ages, but none was made to equate Q-scores. The question is whether the additional experience afforded by doubling the length of the test has any significant effect on quality of execution. Does the pattern of qualitative performance set in the original Maze tend to persist, on the average, in the Extension? If near-equivalence is also found in the qualitative scoring, it would tend greatly to strengthen the evidence for consistency of group behavior in the maze situation, even though individuals vary.

Table 15 sets out the Q-scores, shown, as before, in groups of 50 and 100 consecutively examined, with the S.D.s given for the larger groups.

The agreement between the two series of tests is again very surprising, the largest difference for any group of 50 being a little over 2 points of Q-score, and, if results are tabulated in groups of 100, about 1 point. For all 300 cases the scores were practically equivalent.

TABLE 15

ORIGINAL AND EXTENSION QUALITATIVE DIFFERENCES

N	School	Orig. Mean	S.D.	Ext. Mean	S.D.	Diff.
50	McKinley	22.98		23.04		+0.06
50		26.82		25.24		−1.58
100		24.9	16.3	24.14	16.7	−0.76
	Washington					
50	Leilehua	22.44		24.56		+2.12
50		23.54		23.22		−0.32
100		22.99	14.75	23.89	14.05	+0.9
	Farrington					
50	Roosevelt	22.23		22.22		−0.01
50		25.16		24.76		−0.4
100		23.69	13.85	23.49	14.65	−0.2
300	All Schools	23.86	15.25	23.84	15.37	−0.02

Going back to my 1942 * publication and combining normal (nondelinquent) boys' and girls' averages, I find the mean for 200 cases to be 23.24 points on the original Mazes. This, too, is less than one point away from our present nondelinquent norm. A later study published in 1945 also had 300 cases drawn from high schools in the Territory of Hawaii, and the mean for this group was 21.65 points. In this instance the adult Maze was not used and this would account for the average being lower by 2.21 points.

When sex differences are considered, there is a somewhat changed situation from that found in the mental age scoring. It will be remembered that boys excelled the girls' performance. But in executive quality of performance the situation is changed. The whole group of girls ($N = 150$) had error scores 1.28 points worse than the boys in the Original, but on the Extension they are just about as many points lower, i.e. better, than boys. The differences are, however, not large enough to be significant. Taking the figures as they stand, we can say that while boys are more vigilant in avoiding blind alleys, they do not keep up the quality of their execution when the test is doubled in length.

As regards the comparison of original and Extension Q-scores, the boys as a whole ($N = 150$) had Extension scores 1.23 points worse than the original, while the girls' group was 1.27 points

* Porteus, S. D. *loc. cit.*

better. Whether these norms will approach mainland data gathered on a similar plan is yet to be proved. It may well be that the figures from Hawaii, being related to mixed Oriental, Portuguese, and Hawaiian children, will show considerable differences in qualitative performance from those of mainland American groups, even though the mental ages be very similar. Since the Q-scores reflect disposition and temperament more than ability, this disparity might almost be expected if, as popularly believed, there are ethnic group differences in these respects.

The differences between nondelinquent and delinquent performances in the original and Extension, as reflected in Q-scores, are of course most important. The data are to be found set forth in Table 16. Subjects were girls and boys committed to the Koolau

TABLE 16

NORMAL AND DELINQUENT EXTENSION TEST SCORES

		Normal		Del.			Normal		Del.	
No.	Sex	T. A.	No.	T. A.	Diff.	No.	Q.S.	No.	Q.S.	Diff.
150	M	15.77	50	14.46	−1.31	150	24.45	50	49.9	25.45
150	F	15.08	50	12.33	−2.75	150	23.23	50	57.28	34.05
300	M/F	15.43	100	13.395	−2.03	300	23.84	100	53.59	29.75
300	M/F	15.43	40	14.7	− .73	300	23.84	40	46.3	22.46

NOTE: Column 3 shows mean test age scores for normals; column 5 for delinquents and an adult criminal group (last horizontal line); columns 8 and 10 give similar results in terms of Q-scores.

Boys' Home and the Kawailoa Training School for Girls, two adjoining institutions for delinquents, plus some subjects about to be committed from the Detention Home in Honolulu. A number of the cases had been examined by the Maze within a year or two and in those cases this original test was not repeated, though the instructions for doing the test were repeated for the Extension. The fact that the Extension series was not given immediately in all cases but usually after several months' time could affect the scores. In order to throw some light on the question, the scores of 36 delinquent cases who were given the original Maze and the Extension immediately thereafter were calculated. Their test ages were 14.25 years in the original, 13.96 in the Extension; Q-scores, 43.3 and 47 points. This is not a large group, but taking results as they stand, they lost .29 of a year in Extension mental age and their qualitative scores were 3.7 points worse. In other terms, extended testing results in lower relative performance with delinquents.

It is noteworthy that the delinquent group as a whole is about 2 years below the nondelinquents in test age, the girls making an especially poor showing. When the Q-scores are considered, there is not much change in the picture as compared with the 1942 and 1945 studies, only original Maze scores being, of course, available then. In the first investigation boys averaged 49.35, in the second 48.44, and in the present study the mean is 49.9 points. Delinquent girls averaged 52.9 in 1942, 62.3 in 1945, and the present mean is 57.28. The groups numbered 200, 100, and 100 cases, respectively, with the sexes evenly divided. The 1945 group of girls were mainly juvenile recidivists.

The group of 40 adult criminals included in this table consisted of former inmates of the Waialee Training School for Boys (later removed to Koolau) who have continued in criminal courses and are now serving sentences in Oahu Prison. In spite of the fact that they represent very unpromising recidivists, their average Q-score dropped to 46.3 points, somewhat lower than the mean for the Koolau boys and about 10 points lower than the former criminal average. This drop in score may be attributed to the fact than ten years ago the sample examined consisted of unselected prisoners, many of whom were Filipinos, some of whom were illiterate. The date (1945) should be noted; also that the present group consisted entirely of young men, some having been examined by the original Maze twice previously over a term of years. They averaged slightly less than 25 years in actual age and may be considered more alert than the ordinary prison population. Their test age was relatively high—14.7 years on the Extension. Under martial law in Hawaii young offenders were often summarily sent to prison.

The slight drop in average Q-score may possibly also be accounted for on the theory that increasing maturity has an effect in improving, i.e., diminishing, Q-scores. Support for this view comes from a study conducted by Margaret H. Sanderson [7] who examined children in the public schools of Jacksonville, Illinois. Her subjects consisted of 50 each of eleventh, eighth, and fifth graders, with sexes evenly divided. According to her tabulation, fifth graders averaged (in whole numbers) 38 points, eighth graders 24 points, and eleventh grade children 21 points. Thus, with increasing maturity there was a fairly significant drop in qualitative errors. Sanderson concludes that there is a definite relationship between age or grade and qualitative score, with most significance attached to differences between the fifth and eighth graders and with slight differences in sex performance.

Our own data herein presented concern mainly tenth graders

and are therefore possibly not closely comparable with Sanderson's results. On the original Maze the mean for the Hawaii children was 24, about mid-way between her eighth and eleventh graders.

Another approach to the comparison of delinquent and nondelinquent performance is through the method of matching pairs in each category by chronological levels and Maze test mental ages, and then contrasting the qualitative scores. The matching process, however, could not be carried out completely, as few of the nondelinquents scored as low in mental ages as the delinquents, some of whom could therefore not be appropriately paired. However, pairs were found for 32 delinquent boys and 33 delinquent girls.

The Extension qualitative score for the normal boys was 22.34 points, very close to the general average of 150 boys whose mean was 23.24 points. The delinquent qualitative mean of their delinquent pairs was 43.2 points, about 6 points below the general delinquent mean. However, the difference in Q-scores between normals and delinquents was still 21 points, which is quite significant.

The girls in matched pairs showed a considerably greater disparity. The normal girls scored 25 points (24.41 for the whole standardization group) as against 56 points for the matched delinquents, a difference of 31 points. This difference is of course significant.

In a previous study "critical" scores were described as levels above which 80 per cent of delinquents fell. This was found to be 29 points for boys, 32 points for girls.

In the case of the delinquents examined in the present study, the application of the original Maze took place at such irregular intervals that the results were of dubious significance and are therefore not quoted. In the Extension, 84 per cent of delinquent boys and 86 per cent of girls had Q-scores above these critical levels. I then compared nondelinquent Extension percentages for each successive 50 cases. In each of the three groups 30 per cent of the nondelinquent boys were above the critical level, while the three groups of girls showed respectively 24, 18, and 20 per cent above a critical score of 32, or a mean of 21 per cent for all cases.

It can be taken as a well-attested conclusion that about 25 per cent of nondelinquents overlap with delinquents. As regards critical scores, Docter and Winder, as previously mentioned, have reported 70 per cent of delinquent boys and 30 per cent of nondelinquents above a critical level of 29. The percentage for nondelinquents agrees exactly with that arrived at by me for the Extension series. Of the criminal group, 80 per cent had Extension qualitative scores above 29 points.

As previously mentioned, there is no explanation for this delinquent and nondelinquent overlap. It is abundantly plain that the defects of personality that the Maze reveals do not necessarily issue in delinquent conduct. It may well be that other tests will reveal antisocial indications in this overlapping 30 per cent of delinquents with normal scores. But a word of warning should be directed against relying on tests that analyse replies to questions which are rather directly related to delinquent conduct, even though the difference between delinquents' and nondelinquents' percentages of replies may be great. The question "Have you ever been in trouble with the police?" would naturally differentiate 100 per cent of cases if truthfully answered. It is unfortunately true that the cloven hoof, or the brand of Cain, is not easily apparent.

However, a good step forward toward the understanding of juvenile delinquency will be achieved when offenders can be divided into those who are more specifically victims of circumstances and those with strong proclivities that are likely to issue in crime. Whether this differentiation can be actually furthered by the use of the Maze seems very probable, but validation of this point will require further research.

Chapter References

[1] Tizard, J., "The Porteus Maze Test and Intelligence: A Critical Survey," *Brit. J. Educ. Psychol.*, 21, 172–85, 1951.

[2] Porteus, S. D., "A Survey of Recent Results Obtained with the Porteus Maze Test," *Brit. J. Educ. Psychol.*, 22, 180–88, 1952.

[3] Berry, R. J. A., and Porteus, S. D., "Intelligence and Social Valuation," in *Research Publn. 20* (Vineland, N.J., Training School, 1920).

[4] Porteus, S. D., and Bassett, Dorothy M., *Sex Differences in the Porteus Maze Test* (Training School, Vineland, N.J., Publn. 22, 1920).

[5] Joseph, Alice, and Murray, Veronica F., *Chamorros and Carolinians of Saipan* (Cambridge, Mass., Harvard University Press, 1951).

[6] Burt, Cyril, *Mental and scholastic tests* (London County Council Publication, 1921).

[7] Sanderson, Margaret H., "Performance of Fifth, Eighth and Eleventh Grade Children in the Porteus Qualitative Maze Test," *J. Genet. Psychol.*, 67, September 1945.

CHAPTER IX

Projective-Expressive Aspects of the Maze

The underlying assumption in the Rorschach, as in all other tests of personality, is that the subject's performance is truly representative or consistent. In other words, if he imaginatively responds to color or movement in the ink-blot, this reveals a deeply engraved perceptive habit or characteristic behavior which may be interpreted clinically. This interpretation is difficult to validate adequately.

The use of the descriptive term *projective* is justified on the grounds that an "unstructured" stimulus, to employ present day psychological jargon, forces S to dredge down into his subconsciousness for responses, the theory being that under such circumstances he reveals his personal way of looking at things which may differ from that of other individuals. In other words, he projects his personality traits into his response.

One difficulty with this explanation, apart from the fact that the Rorschach enthusiast would regard it as naïve, is that there is no such thing as an external stimulus that is truly unstructured. Immediately S sees a resemblance between the ink-blot and some object, the stimulus is structured to that extent. The number of popular responses indicates that some blots are more obviously structured than others.

This, however, is a minor point. The important fact is that, in contradistinction to the Rorschach, personality expresses itself in many ways other than by manner or content of perception. This, of course, has been generally understood since man began to observe his fellow men. "As a man thinks and does, so he is" remains a commonsense truism. The statement is just as true and trite if it is reversed to read "As a man is, so he thinks and acts." Undoubtedly keys to character could be found in the way a man walks, talks,

gestures, works, plays, writes and even looks. Probably physical or motor expression is more revealing than perceptive content. But the difficulty with physiognomy, graphology, the Rorschach, and other hypothetical analyses is to reduce the observations to a system so that they can be even crudely quantified. This quantification is the first step toward validation.

We are all aware that the great actor is the one who can subtly fit his own expressive movements, tone of voice, and gestures to the personality he portrays, so that the pattern is recognizable as appropriate by the audience. The characterization must be subtle in laying hold of the nuances of expression; otherwise the acting deteriorates into mimicry. The whole business depends on the fact that there *are* recognizable patterns in behavior. In judging character, most people do as well as the trained psychologist but fail in explaining how they do it. This does not mean that the process is necessarily intuitive. It may depend on attention to minimal signs, even though the observer is unable to list the clues to character which he uses.

Because the details that make up the motor-intellective behavior pattern are mainly subconscious, their expression is truly projective. This fact is illustrated in the saying that the good personality test must have a façade—in other words, S must be unaware of what he is revealing. This condition is met in the projective use of the Maze described in this chapter. The subject's attention is engrossed in the problem of finding his way through a printed labyrinth. He is entirely oblivious of the fact that in executing the task he reveals his own deeply engrained habits of action and, indeed, his attitude to a problem that presents some difficulties in its solving. It must be confessed that until very recently the psychologist applying the test was similarly blind to that realization.

The discovery that Maze test performance demonstrates rather easily recognizable patterns of behavior was a great surprise to me. It came about through a continued study into the effects of routine chlorpromazine medication on Maze Test performance. The original work had been done under my direction with the assistance of John Barclay, the subjects being chronic psychotics, inmates of two "closed" wards at the Territorial Hospital at Kaneohe, Oahu. These studies have been published and are further described and discussed in Chapter 10 of this volume.

In continuation of this investigation, Barclay, under the general direction of Bernard Bloom, attempted to decide how far the effects of ataractic drugs and lobotomy run parallel to each other. The similarity of effects had already been noted by medical observers,

such as Terzain (1952), Aron (1954), and others. Hence, Dr. Bloom decided to apply tests such as the Maze and the Capps Homograph, which had been proved sensitive to psychosurgery, and also a selection of tests such as the Wechsler, which had shown relatively insignificant changes after brain operations.

In previous investigations, notably the Columbia-Greystone, the original or standard Maze series had been applied both before and after operation without regard to practice effects. Unfortunately, in this new psychopharmacological study, the original Maze was also used after medication. Thus improvement through practice diminished any deficits due to the use of the drug. Hence, the results were no check on our previous investigations in which we controlled practice by using the Extension Maze series after medication. However, it was reasonable to assume that if patients showed declines in score when the original Maze series was reapplied, the real deficit would be the amount of decline plus the average practice improvement.

In looking over the patients' records, I was struck with the general resemblance of execution of the two series of tests by the same patient, even though about three months had elapsed between the first and second examinations. For some patients the two performances were so alike that it seemed possible that they could be identified as the work of the same persons, even if the names were hidden.

To test this supposition, we concealed names and all other identifying data by attaching strips of cardboard across the face sheets. We found that it was possible to match correctly the two Maze performances in about 60 per cent of cases. Furthermore, we noted that attention to such details as beginnings and endings was of the greatest assistance in the matching process, although other more general features, such as pencil pressures, types of corner turning, and over-all similarities were also helpful in doubtful decisions. In some instances increased tremor made the tests more difficult to match. This condition is not rare with chlorpromazine patients.

By actual experiment we found that the mean percentage of correct matchings at random in fifty trials, using 10 pairs of playing cards, was actually 8 per cent, and this result was so different from the 60 per cent correctly identified by inspection of records that we determined to continue the experiment, using inmates of the schools for delinquent children as subjects.*

Most of these inmates had been tested at some time or another

* This study was undertaken as the preliminary portion of a psychopharmacological project supported by funds from the National Institute of Mental Health, Bethesda, Md., 1958–59.

by the standard Maze so the full Extension series was applied. Because some subjects had not been given the adult test on account of succession of failures at lower levels, we decided to repeat for matching purposes only three of the easier mazes, the VII, X, and XI. In order to make recollection of previous performance less likely, the first five tests of the Extension series were given in the first examination, followed immediately by the repeated applications of the VII-, X-, and XI-year tests. The VIII- and IX-year records were not used in matching. The actual process consists of first clipping together each individual's Maze records and marking them plainly with the numbers 1 and 2 to distinguish them as first or second applications. The number ones (with the VIII and IX designs removed) are then spread out on a table. For convenience of inspection and comparison, the records are matched in groups of tens. The examiner then picks up the number two records and compares them one at a time with each of the Year-VII mazes exposed on the table. Those that most resemble the specimen to be identified are segregated and the comparison is continued using the X- and XI-year records. On the basis of his final judgment about which two were done by the same person, the examiner then lifts the strips that conceal the names. If the two are matched correctly, both are withdrawn from further consideration and scored as one pair correctly matched. If the judgment is incorrect, the number two record is put on one side, the number one specimen being put back on the table so that it remains as a possible match with the rest of the number twos. Obviously the other half of the correct pair must be still included in the as yet unmatched pile.

A variant procedure which often saves time in matching is for the examiner to go rapidly through the pile of number two records and pick out any with some marked characteristic, such as right-angled turns, and attempt to match these first with those exposed on the table.

It was soon apparent that ease of identification differed greatly in individual cases. In some instances a judgment could be made on comparison of the VII-year record but verified thereafter by comparison with the other two tests. On the other hand, there are cases in which all the records must be carefully considered in detail and a very fine balance in judgment is required. In such instances the resemblances are so slight that the matcher is very doubtful as to his final decision.

Following up our previous experience with the psychotics, we found it possible to limit the comparison of performance details to the beginnings and endings of the maze-tracings, with only very

infrequent reference to more general resemblances. The percentages of correct matchings (in groups of ten individual records each), rose to rather extraordinary levels. The summary of results is shown in Table 17.

TABLE 17

PERCENTAGES OF CORRECT MATCHINGS

Group	Porteus	Barclay	Mean	No. Matchings *
Nondelinquent Boys—Kam. School N = 50	90.0	93.0	91.5	11
Nondelinquent Boys—McKinley N = 50	73.3	96.2	84.8	12
Nondelinquent Girls—Kam. School N = 50		91.0	91.0	10
Delinquent Boys N = 60	90.0	96.7	93.3	18
Delinquent Girls N = 50		96.7	96.7	6
Oahu Prison—Orientation Males N = 40	96.2	90.0	93.1	12
Oahu Prison—Recidivist Males N = 40	90.0	92.5	91.2	12

* Column 5 gives the total number of matchings of the same groups upon which the mean in the preceding column is based. Thus it actually shows the number of times the experiment was repeated.

It will be observed that the average percentages of correct matchings (column 5) range in round numbers from 97 to 85 per cent. The lower figure is affected by a wide difference in accuracy of matchings between the two examiners in the case of nondelinquent boys. Here my own average of 6 matchings was only 73 as against Barclay's 96 per cent. The discrepancy may be accounted for largely by the fact that my matchings represented the first attempts with nondelinquents and before it was fully apparent that beginnings and endings provided the most significant clues to identification.

Barclay's matchings were done much later after the system had been fully developed. When this stage had been reached, his matchings and my own never varied more than 10 per cent on any individual groups. Limitations of time prevented more work being done in this field but in every experiment with individual groups of ten Ss, matchings were repeated at least twice. Obviously, when any observer matched 100 per cent of a group correctly on two occasions, there was no point in proceeding further.

The evidence presented shows that after some experience, matchers can achieve successful identification of records in 90 per cent of instances, no matter whether the groups are delinquent or nondelinquent. This means that all individuals tend to set up recognizable patterns of response on the basis of a single application of the Maze. Furthermore, the strength of the tendency to conform to that pattern differs in individuals. How much it differs and the meaning of the difference in terms of social behavior are obviously questions that call for answers through extended research. The fact that only beginnings and endings are necessary for matching directs attention to an additional problem: What is the minimum performance necessary to reveal individual characteristics of response? This matter will be considered later.

It had already been apparent in many cases that the interval between test applications did not have very much bearing on matchability. In the great majority of instances, the second application took place immediately following the first, but in some cases, notably recidivists at the prison, an interval of months had elapsed between testings. Yet the performances could be easily matched. The influence of lapse of time on matchability was made the next subject of enquiry.

It was found that after an interval of about three months, between 70 and 80 per cent of delinquent performances could be matched but the percentage of identification with normals fell to between 30 and 40 per cent. This result warrants the conclusion that as a group delinquents are less flexible; in other words, they find it much more difficult to break an established pattern of behavior.

The next problem concerned investigating the possibility of scoring the resemblances so that a numerical "index of matchability" would be available. The similarities that serve as clues for identification may be briefly mentioned. In tests VII and X the starting point is indicated by an arrow, in Year XI by an S. Where the subject begins his maze tracing in relation to the arrow (or S) is a significant clue. Some begin exactly at the arrow point or on the shaft but more commonly somewhat short of the point. The tend-

ency is to start the second application at approximately the same place.

The final exit is marked by an arrow in Year VII and S may end short of the arrow shaft or may go through, under, or over the arrow. Ending beyond the arrow is very symptomatic. In Years X and XI type of exit is important. Some Ss prematurely consider the maze solved and do not pass through the exit at all. Others finish the task with a flourish and in some cases end the stroke with a very fine line. This we call "feathering," and it is probably indicative of S's attitude to the maze as a whole. After a cautious and even meticulous performance, the end is in sight and the final flourish with relaxation of pencil pressure may be an expression of satisfaction that the task is over and done with. This interpretation of attitudes is being investigated fully. If verified, it may give the psychologist, at least in this area, a power with which he is sometimes credited but does not possess except for emotional situations, namely, a reasonable certitude as to the mental attitude of an individual in approaching and carrying out a task.

To arrive at a similarities score that is intended to serve as an index of matchability, we have prepared a check list of items of resemblance in two performances in the one test. They are in six divisions, one for each beginning and each ending in each of the three tests. Pairs of similarities in relation to the arrow and exit, if S began or ended his second tracing within $\frac{1}{16}$ of an inch of his former effort, and the paired resemblances of the first and last turns in the mazes are recorded on the check list by plus signs. Actually this scoring does not entail comparisons of more than 1½ inches of each tracing and often less, except in one instance. This is the beginning of the XI-year design in which general similarity is judged by inspection of the records up to the third turn. General similarity may involve such comparisons as pencil pressure, "feathering," and irregularity of line, which would lead the examiner to conclude that the same person did the two mazes under inspection.

The number of plus marks gives the consistency or similarities score. Generally speaking, the size of this score corresponds with the ease or difficulty that would be experienced in matching. It is not expected that many psychologists will go through the rather tedious process of matching. The consistency score will be all that is necessary. The check list follows. Though not tabulated in the present volume, the range of consistency scores is from 4 to 21 points, the averages being 13.2 for nondelinquent girls, 11.5 for boys. Apparently the latter are less rigidly consistent. Delinquent boys, on the other hand, score 14.6 points.

SIMILARITIES CHECK LIST

YEAR	BEGINNINGS		ENDINGS	
VII	Both short of arrow point " on " " " below " " Both above arrow point " before " " Begin within 1/16" of each other First turn very similar Draws alike to second turn		Both short of shaft " on " Both below arrow shaft " above " " Both past arrow point End within 1/16" of each other Last turn very similar Draws alike from second last turn	
X	Both short of arrow point " on " " " through " " Both left of arrow shaft " right " " Begin within 1/16" of each other Draws alike around first turn		Both end short of exit Both go through exit End within 1/16" of each other Turn at exit alike Last 1½ inch alike	
XI	Both short of S (start) Both on S Both level with or below first opening Begin within 1/16" of each other First turn similar Draws alike to third turn		Both end short of exit Both end through exit End within 1/16" of each other Turn at exit alike Last 1½ inches alike	

Method: After two applications of the Extension series have been assembled in pairs, check marks, usually plus signs, are placed opposite the item in which both tests agree. As an example, if the points of beginnings and ends of lines (otherwise starts and finishes) are within 1/16 of an inch of each other, those resemblances would each rate a plus. Similar angles or curves at the first and last turns also earn plus marks. The total number of check marks is the similarity score. High scores signify consistent performances, low scores variable reactions.

The method of arriving at a consistency score was adapted only after considerable experiment. At first, numerical values were assigned according to the degree of resemblance. For example, in scoring the position where S began the maze-tracing in relation

to the arrow on the Maze blank 3 points were given if the point of beginning was exactly the same in the first and second performances, 2 points if $\frac{1}{16}$ of an inch or less difference existed, 1 point if between $\frac{1}{16}$ and $\frac{1}{8}$ of an inch separated the two beginnings or endings, and 0 if the distance were greater than $\frac{1}{8}$ of an inch. The angle at the first turn could be similarly scored 3 for a right angle, 2 for a near right angle, 1 for any angle, and 0 for no angle at all. But, on the other hand, curves could be exactly similar and score nothing. In making judgments objective, we often lost sight of over-all resemblances. We decided then to sacrifice detailed comparisons in favor of a check list, which might be less numerical in nature but still be truer, on the principle that a picture, no matter if somewhat impressionist, might be more recognizable than a line drawing.

In discussing the results with Barclay (who has been very closely associated with me in all the later development of the work mentioned in this chapter), we wondered what percentage could be matched if only a very limited amount of the subject's tracing were exposed for the purpose of comparison. Barclay then took a group of ten nondelinquent girls' records and, after cutting "windows" in blank designs, showing only the beginnings and endings to the extent of not more than an inch, he fastened the blanks over the girls' tests. Thus only a very small portion, in the case of beginnings not even the first turn, was exposed through the window, the rest of the tracing being concealed by the blank design. As regards endings, the window was only a half inch in width with a maximum length of an inch. The object was to determine the relative importance for matching of the initial and finishing strokes of the pencil.

To my great surprise, Barclay was able to match correctly 80 per cent of records instead of the 30 per cent which we expected. I also matched 80 per cent and Henry Culver, research assistant, correctly identified 90 per cent. He used the procedure of a quick survey of the pile of second performances to pick out any with distinctive features, such as right-angle turns, for easy matching.

What this means is that it is necessary to inspect only a tiny sample of a person's complete performance in order to pair it with his second performance. By actual measurement, the proportion that beginnings and endings exposed in this way make up of an individual's total performance in Years VII, X, and XI is less than 3½ per cent. If the length of his tracings in all eight tests of the Extension series is measured, it makes up less than one per cent of his complete drawing, and this assumes that he threads all mazes

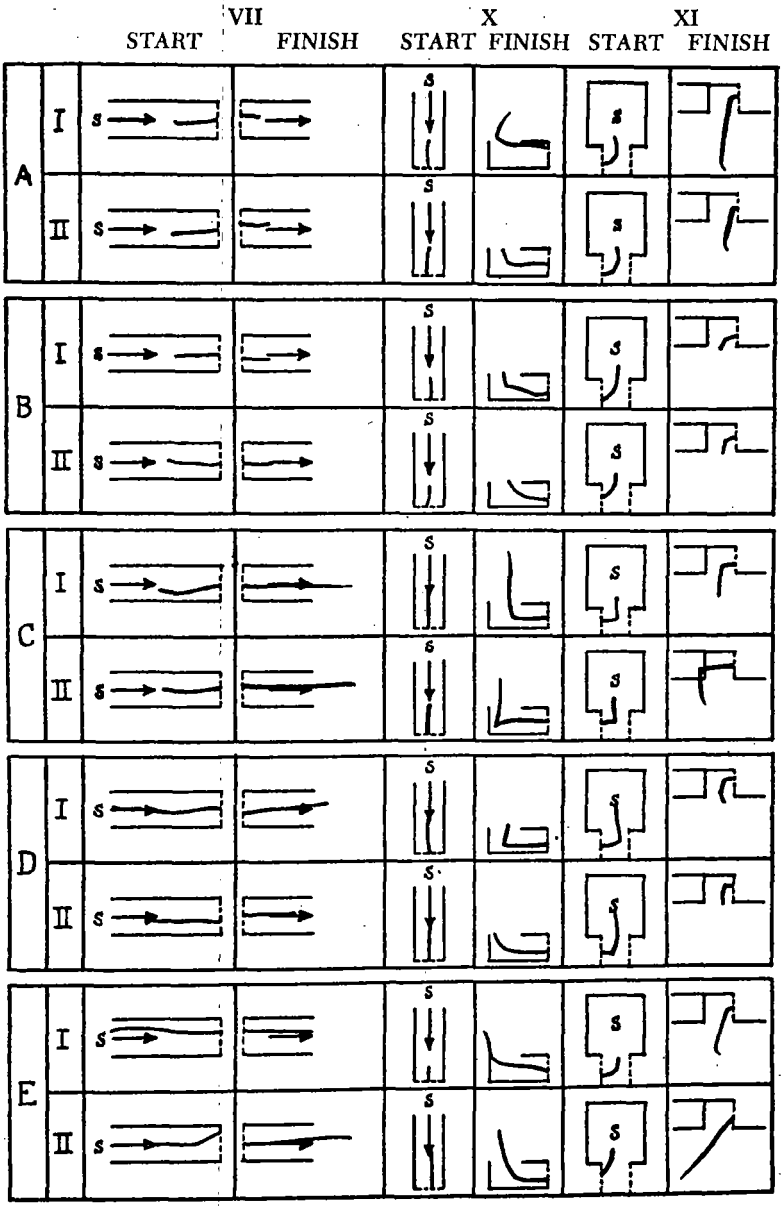


FIG. 2. Beginnings and Endings

on one trial each. In other words, it has been proved possible to establish an individual pattern of behavior in 80 per cent of individuals on the first and last half to three-quarters of an inch they draw in tracing each Maze. This could only be done if this sample were truly representative. Figure 2 gives the paired performances of five individuals (designated *A, B, C, D,* and *E*), showing beginnings and endings, the first and second tracing being numbered I and II, respectively. Only the parts of the tracings exposed to view through the windows appear. Comparisons can be made of tracings in pairs shown in the vertical columns. It will be seen that beginnings are most similar. Individual *E* was very difficult to match.

That patterns of expressive behavior are so quickly set up as to make the *second* threading of a maze design recognizable as being done by the same individual seems to the writer to be a very extraordinary phenomenon. The crux of the matter appears to be the subject's preoccupation with finding his way out of the maze, which absorbs his conscious attention, allowing his subconscious tendencies to reveal themselves. How persistent is the expression of these tendencies is illustrated by reproducing in Figure 3 two performances by a delinquent boy on Year X twelve months apart. Special attention is directed to the beginnings and endings of the maze tracings as well as to the over-all resemblance. Figures 4 and 5 show paired performances on Year XI, done by a normal girl. Figures 6 and 7 show similar performances by a delinquent girl on the same test. They serve to illustrate the contrast in the general execution by the two types of cases, and at the same time the similarities of beginnings and endings done by the same individual upon which matching is based.

Similarities in signatures or writing in general are, of course, commonplace in their appearance, but they naturally are expected when countless repetitions of the same letters are involved. Again it must be emphasized that the particular maze-threading operations under review have been repeated only twice; consequently, there is no presumption that reactions have become automatic. The question about what influences the subject to make his first and last strokes in a maze-tracing, either in a similar position or with the same kind of turn, is puzzling. All that we can suggest is that we all tend to develop a personal attitude toward beginning and ending a job and that we give motor expression in some way to that attitude. No doubt the movements of a business executive before he begins to dictate an important letter, if they could be recorded without his knowledge, would reveal similar set patterns.

It would be tempting, after the fashion of the psychoanalysts,

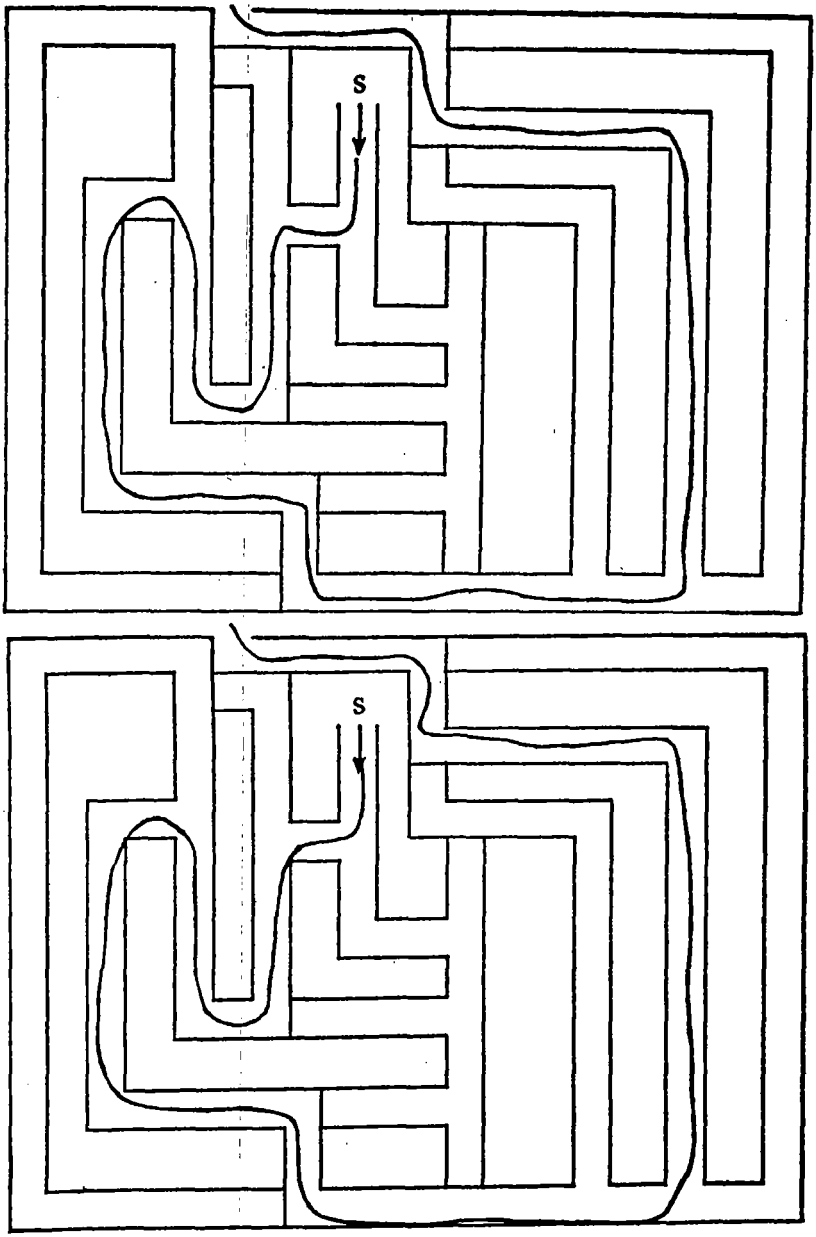


FIG. 3. Matched Performances of a Delinquent (One year apart)

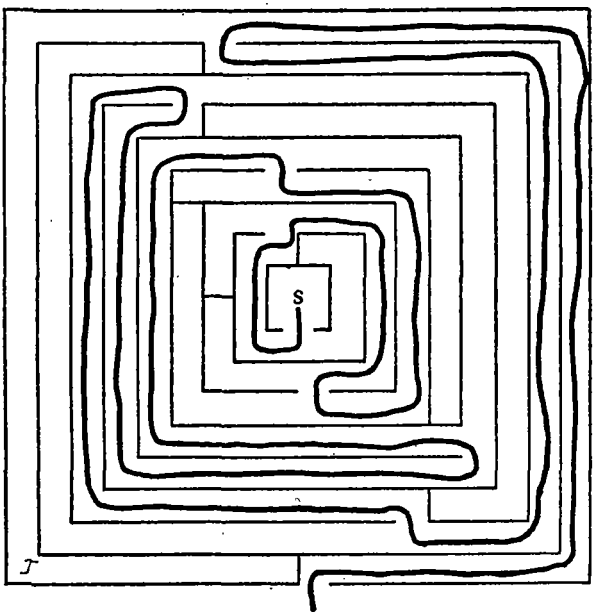


FIG. 4. First Performance Year XI
(Nondelinquent girl)

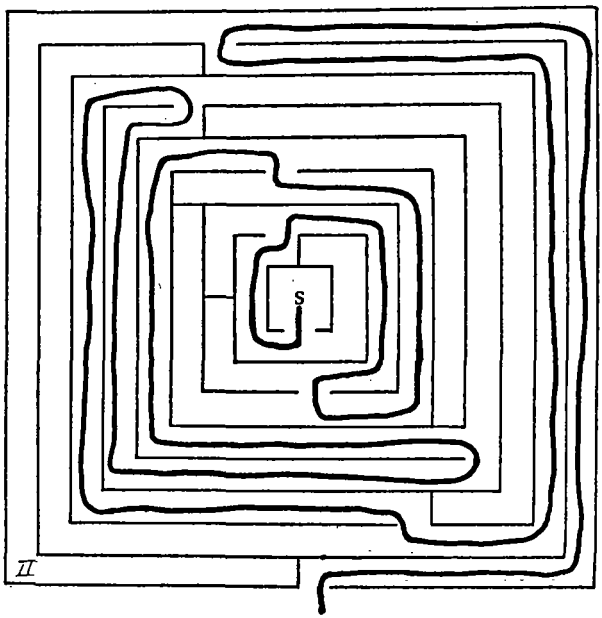


FIG. 5. Second Performance Year XI
(Nondelinquent girl)

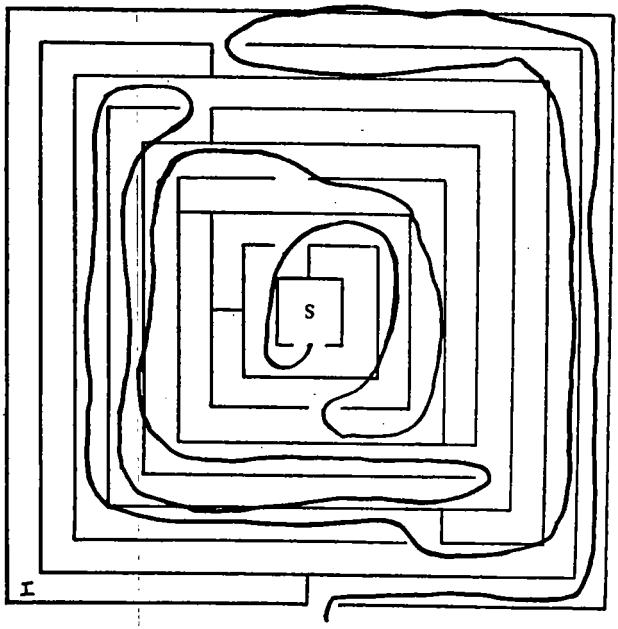


FIG. 6. First Performance Year XI
(Delinquent girl)

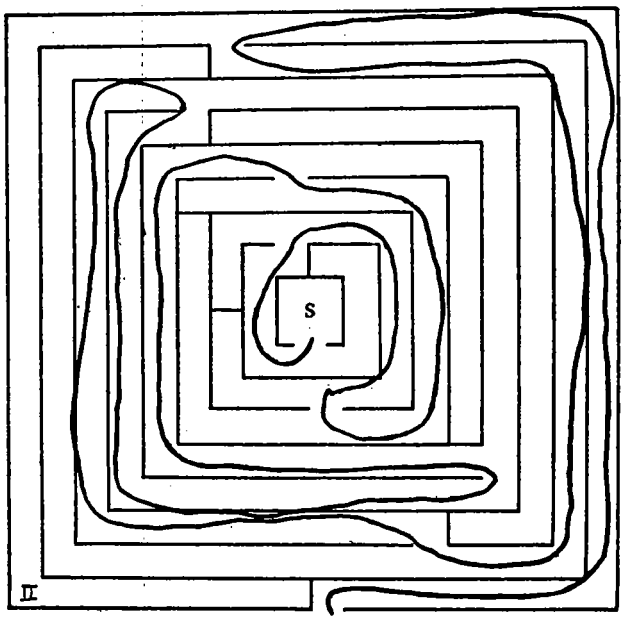


FIG. 7. Second Performance Year XI
(Delinquent girl)

Rorschach devotees, and graphologists, to consider similarities as projecting definite traits of personality of the subject. We have already suggested an interpretation of "feathering" as an expression of satisfaction with the successful completion of a trying task. One might possibly interpret short beginnings as expressing the subject's desire to get the task begun and finished as quickly as possible, thus arguing an impatient attitude toward work. The task of validating such assumptions would be laborious, involving correlating speed of performance with the tendency to begin "short." Right-angle turns also may be interpreted as indicating a tendency toward neat, precise execution, but one of the first questions to be answered concerns the frequency of its occurrence with different types of subjects.

At the beginning of this matching experiment I was struck with the number of right angles occurring in the performance of recidivists. Later, I discovered that my suggestion that this kind of performance was characteristic of compulsive behavior had been anticipated by Foulds in England seven years ago in his observations on maze-tracing by obsessional neurotics. His findings have already been quoted.

Concisely stated, it may be that calling these details of execution expressive-projective aspects merely raises the further question —expressive or projective of what? Before there can be any adequate answers there must be a tremendous expenditure of research. This also suggests that there will be a shift of interest away from generalized results of Maze application, such as mental age, toward careful consideration of details of performance. Landis and Erlick have already provided a lead in their classification of kinds of errors in the Maze, particularly those that indicate mental confusion. Even such gross analysis should help greatly in studying the behavior of abnormal types.

Incidentally, my discovery of the frequency of right-angle turns among habitual criminals' performances changed my clinical interpretation of this feature. Formerly, I had considered it characteristic of the careful, neat, dependable, reliable, methodical worker. This may be true, but if so, it is just as good an indication of the consistent tendency of the individual to do wrong as to do right. I should like very much to pit fifty bishops against my fifty recidivists and compare the incidence of "compulsive" behavior. The performance of meticulous housewives might also be illuminating.

Having determined that patterns of beginnings and endings in Maze performances are characteristic of individual performance, the next step was to attempt to discover the meaning of such pat-

terns. To simplify the problem by limiting still further the data to be examined, beginnings only were studied in delinquent and nondelinquent performance. This involved counting and comparing the total number of times individuals began short of the designated starting points in contrast to the occasions when they began actually tracing on those points. The difference could then be expressed as a ratio of incidence.

But in examining the records, it was noticed that subjects sometimes started carefully on the mark in the easier tests, but as soon as difficulty was experienced, their tendency to start short was more frequent. Hence, the ratios of "on" to "short" beginnings in tests XII and above were also calculated. For this part we considered the full record of each individual in the Extension series from Year VII to the Adult test. In order to diminish any possible relation of racial temperament to style of response, all cases considered were of part-Hawaiian blood, divided into nondelinquents and delinquents.

Consideration of the results of these comparisons, divided by sex, again reveals a remarkable situation. For the whole series of tests, nondelinquent girls had double the ratio of "short" to "on" beginnings; for tests XII and above, the ratio for normals was 1.5 to 1 but it rose to 5.13 to 1 for delinquent girls. The frequency of this beginning pattern for the latter group is extraordinary and calls for theoretical explanation.

Since the tendency increases so markedly when the tests become more difficult, frequently after a failure in a test trial has occurred, the phenomenon would seem to reflect the temperamental attitude to a puzzling situation. Does it mean that the delinquent girl is unwilling to face and solve problems, or is she eager to get the disagreeable task over with? Data on the relation between speed and short beginnings would be helpful in answering these questions and these are being collected and studied.

However, the sex differences would indicate that boys deal with the Maze in another fashion. Their tendency to react to the more difficult maze designs does not change so markedly as the tests become harder.

One fact, however, seems of great importance when sex performances are compared. Over 90 per cent of the girls committed have been involved in sex offenses; less than 10 per cent of the boys have been committed for such offenses. If further enquiry confirms the relationship of short beginnings to sex offenses, then for the first time the type of delinquency could be related to an objective test response.

Fortunately, "shortness" can be expressed exactly by measuring with dividers the distances between the actual beginning and the point of the arrow or the letter S in the middle of the extension design. The tendency of delinquents to start short is again demonstrated. For 47 delinquent girls, the mean distance from the designated point of beginning was 33.9 mm, while for 40 nondelinquent girls it was only 15.3 mm, or a ratio of 1 to 2.2. This disproportion again demands explanation. Does this disregard of the rules laid down for conduct in tracing a maze indicate S's attitude toward the rules of society, especially in sex behavior? All that we are justified in presenting at this time is the simple affirmation that nondelinquent part-Hawaiian girls begin difficult mazes differently from delinquent girls of the same racial origin. Furthermore, the ratio of "short" to "on" beginnings is 6.01 to 1 for all cases of delinquent girls as compared to 1.15 to 1 for nondelinquents. Apparently the only way to interpret such extraordinary differences is by devising individual rating scales of behavior.

As matters stand at present, the success of the matching process has a bearing on the reliability of the qualitative score in as far as that score reflects consistency of performance. In fact, no test-retest correlation can be more decisive than the demonstration that two performances of an individual can be paired by inspection, since that fact proves consistency of reaction to the test.

The applicability of the reliability coefficient as obtained from test-retest data is dubious in respect to the Maze. Some psychologists have not realized that the nature of the test, plus the method of allowing repeated trials in certain tests, introduces the factor of learning or adaptation within the limits of the test itself. Since individuals vary so greatly in their ability to learn from experience and since a high coefficient of reliability reflects consistency or sameness of response, the score cannot possibly measure these opposite tendencies.

Actually the wide discrepancies in results achieved by different investigators should have made the application of the "reliability" formula suspect. Dr. Clement, of the University of Paris,* obtained a coefficient with mental defectives of .86; Docter (1957), working with psychotic patients, found it to be .55; Morgenthau (1922), presumably using a split-half method, obtained a coefficient of .95; Mackie (1929) found it to be only .29 with normal children. Which of these figures is to be relied upon? My own answer to the question would be that they are probably all irrelevant.

Undoubtedly, a split-half method would be the fairest, but how

* Private communication.

should the test series be split? Picking alternative tests would hardly achieve an equal division. Dr. Milton Jensen has suggested what would seem to me to be the only means of dividing the test so that each half would adequately represent equal learning opportunities within the test series. He has proposed numbering consecutively all choice points where mistakes are ever made in maze threading, then counting an individual's scores on odd and even numbered choices. He believes that this would yield a very high coefficient of reliability. This would apply to the test-age score. The fact that, in the great majority of instances, changes in score are in the one direction, namely improvement, would surely indicate that such changes are due to practice. In other words, differences in score on a retest reflect the variability of human subjects, not of the instrument. With regard to the qualitative score, the matchability of the same individual's repeated performances offers a new but convincing proof of reliability.

It would be tempting indeed to offer a discussion of the implications, educational and otherwise, of this new dimension in Maze testing. If, for example, it could be confirmed that delinquents are much more set in their patterns of behavior than normals, it should prompt reconsideration of the methods used in training such cases. Surely, strict regimentation, fixed discipline and imposed regularity in personal activities would then be tantamount to gilding the lily, or impressing the mold of habit on individuals already too susceptible. Even the rival philosophies of progressive and conventional education of normal children might possibly be affected in view of the fact that inflexibility of reaction is characteristic of the vast majority of our cases, whether delinquent or nondelinquent.

Exploring such avenues as the meaning of types of Maze reactions, no matter how wide they may be, should await further research. We need to know how characteristic responses are affected by age, sex, race, social adjustability, education, religion, mental abnormality, and the like. In addition to this study of group differences, there is the far more exacting task of finding out the meaning of individual differences in similarities scores, number and extent of short beginnings, changes in tendencies within the one series, or after more extended testing. The problem of interpretation is difficult, because as yet we have no means available of adequately quantifying social behavior as a measure for use as an external criterion.

In any case, the opening up of these new dimensions in Maze test application may easily entail ten years of intensive effort before the practical application of the findings is reasonably certain. How-

ever, when we consider the immense amount of research that has been devoted to the visual interpretation of ink blots or, earlier still, to the interpretation of Binet scores, it would appear that the Maze has more than enough clinical promise to make research on its application worth while. Some of the reactions briefly described in this chapter are so amazing that I have every confidence that the challenge to clinical interpretation, which they present, will be eagerly accepted. For my own part, I find these projective-expressive features more interesting than standard test scores.

CHAPTER X

Maze Deficits After Chlorpromazine

Evidence has been presented in previous chapters that serious impairment of Maze-tested functions follows various types of psychosurgery, including lobotomy, ablation of various areas of frontal lobe cortex (topectomy), interference with blood circulation of these same areas (venous ligation), surgical damage to the thalamus, and partial removal of the temporal lobes of the brain. The fact that Maze defects are so consistently evident after frontal lobe damage suggests that planfulness is at least partially mediated by the superior prefrontal portions of the brain. Moreover, the association is comparatively unique as far as mental tests go. However, ability in maze-threading depends upon conditions other than intact frontal lobes.

One small but extremely significant study demonstrated that similar deficits were associated by experimentally induced vitamin-B deficiency, the losses increasing as the deficiency deepened and being recovered after thiamine or B-complex feeding. Again, other tests showed no such results.

In a very early experimental study carried out in North Queensland Waite and Neilson [1] showed that children heavily infested with hookworm attained lower Maze scores than the lightly infested children, but hookworm-free cases showed no losses. They reported their results in 1919. Other debilitating diseases such as pellagra and beriberi could be expected to be followed by Maze impairment. The association of test performance with hypoglycemia and diabetic conditions should be examined.

If additional proof were needed that the Maze is a sensitive psychosomatic measure, reflecting not only brain damage but also other conditions affecting the nervous system through blood chemistry, it is almost spectacularly provided by studies of psychotic patients under routine dosage of chlorpromazine, one of the most effective tranquilizing drugs.

Psychiatric opinions based on hospital experience seem to agree very definitely that many psychotic patients do show changes in behavior or personality after fairly prolonged routine administration of the so-called ataractic drugs. A common or medium dosage in the case of chlorpromazine would be 300 mg. a day. But the drug does not always act as a tranquilizer, especially in the case of some chronic psychotics. Occasionally there is a stimulating influence observable in mute or stuporous patients. After several weeks of medication at one of the closed wards of the Territorial Hospital at Kaneohe, Hawaii, the psychiatric aides noticed an increase of minor injuries. Often elderly and infirm patients, who had been content to remain inert all day, were stimulated to voluntary movements about the ward. Their steadiness and coordination were inadequate because of long muscular disuse and many fell. Undoubtedly the drug itself contributed to their physical incoordination.

Uncooperative attitudes were ameliorated to the extent that patients took a more active part in ward work, and some were able to be scheduled to outside light labor. To call these effects "ataractic" is somewhat misleading. Nevertheless, the most dramatic changes were in the direction of allaying anxiety and the diminution of aggressive, noisy, or destructive behavior. In some cases, after an initial sharpening of hallucinatory and delusional phenomena, these aspects of psychotic behavior became less apparent; "drive" decreased and pressure of activity in various directions fell to more normal levels. In speech, for example, the extremely talkative talked less, but the mute and withdrawn cases became, to some degree, vocally more expressive. Under such circumstances it would seem much better to describe the new psychopharmacological aids as stabilizers rather than tranquilizers.

With the aid of a ward behavior-rating scale, specially devised for application by psychiatric aides trained in giving ratings, I [2, 3] undertook to measure changes in the psychotic personality following medication. The traits covered by the scale were aggressive or destructive behavior, negativism or lack of cooperation with others, speech disorders such as either extreme loquacity or under-vocalization, untidiness, restlessness or physical overactivity, depression, hallucinations, delusions, mental confusion, asocialization, and compulsive behavior.

In case others should use such scales, a word of caution should be given. First among defects in their application is the failure of untrained raters to realize that except with regard to hallucinations, delusions, and possibly compulsive behavior, ratings in the

scale are intended to cover normal as well as psychotic reactions. Consequently, the ratings in many instances do not show a normal spread. I found it necessary during the training period to have aides rate themselves and each other. I was able to show that my own self-ratings would indicate by no means zero ratings on aggressiveness, talkativeness, physical restlessness, while others might consider me moderately untidy, uncooperative, or asocial. Part of the difficulty arises from the special meanings attached in hospital use to some of these terms, as, for example, "untidy." As a result of these demonstrations, a much more normal distribution of ratings on patients was obtainable. Only psychiatric aides who showed ability to justify their ratings on individuals by specific evidence and who were fairly self-consistent in repeated judgments, were considered as possessing an objective attitude as well as an understanding of the meaning of terms used and were selected as raters. The cause of research in hospitals would be greatly furthered if the nature and application of behavior scales were fully covered in training courses for personnel.

Other weaknesses of the scale are those met with by devisers of all such rating devices, in addition to those specifically related to measuring the therapeutic effect of any new discovery such as a "tranquilizing" drug. One of these is suggestibility both of the patient and of the individual who is attempting to observe or measure effects. Any drug that a doctor prescribes with a statement that it is something new and that he thinks it will do the patient good raises an expectancy of improvement. The psychiatric aide who first observes the results may be skeptical, but when he finds that real changes in behavior are quick to occur after use of the drug, his doubts may change to overenthusiastic acceptance. This increases markedly after patients who have, in spite of all other treatment failed to respond, now seem to change and admit a sense of well-being.

Ordinary observations of patients seem to indicate that a drug like chlorpromazine, after an initial period of "dopiness," exerts a marked initial impact on behavior. Appetite, sleep, spirits, all improve. During this time the level of confidence and expectancy rises in both patient and rater, reflected by quite noticeably higher ratings.

Unfortunately, bodily habituation, like mental progress, is rarely consistently steady, but is subject to plateaus. When a patient's advancement levels off, he is disappointed, and this is reflected in lowered ratings. Even the best raters will exhibit ups and downs, partly owing to the variability of psychotic behavior, and partly because these fluctuations in effect overinfluence judgments.

When ratings are marked on a linear scale, numerical values can

be assigned to each, and the whole can be summated as a score before and after medication. The summation presents another difficulty, one that besets the path of any psychologist who combines a number of discrete items, or subtests into a scale. The contributions of each subtest to the total score obviously depend on the comparative value or validity of each.

Thus in the Stanford-Binet each item, if successfully dealt with, is assigned a two months' credit as regards the subject's mental age. A ten-year-old child earns exactly the same credit in Form L of the Terman-Merrill Binet if he can remember the items of information in a news story about a fire that he does if he can give 28 words in free association in one minute. The same credit accrues if he can repeat correctly six digits given in nonconsecutive order by the examiner. According to its author, the Wechsler-Bellevue Scale is also predicated on the idea that all the subtests have equal value. The Arthur Point Scale, on the other hand, attempts to weight its component subtests, but on a criterion that has little or no relation to their comparative validity. This very doubtful assumption of equality of value is also made in a behavior scale that adds together in a final score numerical ratings in discrete traits or trait complexes.

These uncertainties of method and application notwithstanding, the Ward Behavior Scale may be accepted as an approximate measure of personality change after chlorpromazine. The average of two premedication ratings was taken as a baseline in reckoning change after the drug, and six postmedication ratings were averaged in pairs—the 3- and 6-, the 9- and 12-, and the 15- and 18-week scores. Hence, there were 4 points to be established on a graph. This process of pairing ratings tended to smooth out fluctuations. Differences of 9 or less points from the premedication scores were disregarded as slight or negligible, while changes from 10 to 14 points in the scale were considered to represent moderate improvement, 15 to 19 points marked improvement, while a shift of over 20 points was considered to mean very marked change.

To disregard the changes in ratings below 9 points constitutes a very considerable allowance for the various weaknesses of the scale. But as will be seen from Table 18, the control group to whom placebos had been administered under the so-called "double-blind" procedure showed significant losses in only 5.5 per cent of cases, while 11.3 per cent gained. The experimental group showed no significant declines, but 60 per cent showed significant gains. It would seem fair to conclude that about 50 per cent of chronic psychotic patients improved in behavior over and above the per-

centage of placebo cases in which gains might be ascribed either to suggestibility or to those changes which take place naturally without benefit of ataractics.

TABLE 18

WARD BEHAVIOR CHANGES SUMMARIZED

(N = 50 in each group)

Rating Point Changes	Controls		Chlorpromazine Group	
	Worse	Better	Worse	Better
0 to 4 (Negligible)	16.6%	16.6%	4.0%	12.0%
5 to 9 (Slight)	5.5%	44.5%		24.0%
10 to 14 (Moderate)	5.5%	8.3%		26.0%
15 to 19 (Marked)		3.0%		26.0%
20 plus (Very marked)				8.0%
Above 9 (Significant)	5.5%	11.3%		60.0%

I am inclined very much to doubt whether in the type of ward patient we worked with there would be, under ordinary hospital custodial treatment, any such significant improvement in four months. Because of the extreme paucity of staff—there were only two aides on duty daily for 80 patients, with occasional visits by a staff psychiatrist—no other therapeutic measures were possible. To expect to find 10 per cent of the chronic patients making noticeable improvement under such circumstances would be most optimistic, leading to the belief that all that was necessary with psychotics would be to shut them up in large groups and in time a great number would recover. This, of course, is a vain hope. As matters stood, the treatment afforded was similar to what must be standard in many overcrowded, understaffed state hospitals. This bankruptcy of routine psychiatric ideas and practices was, in many patients, clearly exemplified by the case histories. Several had had lobotomy operations, many had suffered insulin shock treatment, while the record number of electric shock treatments for any individual was 257.

Probably the improvement noted in placebo cases should be ascribed to suggestibility, both group and individual. Under group effects I would include the sudden change from a custodial atmosphere to that of a real hospital, in which drug therapy was in action and patients were improving, resulting in favorable environmental changes, such as lessening of noise, decline in assaultive

incidents, and other disturbances. It is no exaggeration to say that hope entered through the bars of the closed ward with the chlorpromazine tablets. There is also good psychological foundation for believing that the administration of placebos themselves is in part a therapeutic procedure, but until groups can be equated for suggestibility, I cannot see that the placebo device, no matter how much it is in favor with psychiatric researchers, has much value as a control measure.

Rosenthal and Frank [4] have pointed out that improvement following the administration of a placebo is a common phenomenon in a variety of situations. In other words, the use of a placebo does not eliminate suggestibility, and this raises grave doubt about the value of the procedure as a controlling device. They cite a study on the effect of a vaccine on the incidence of common colds. Over a period of time 55 per cent of subjects reported a reduction in the number of colds after vaccine, but 61 per cent claimed similar increased resistance after placebo. Their own study showed that not only were favorable symptoms reported after placebo, but suggestion also brought about appearance of unfavorable conditions. Sixty-nine per cent had decreased blood pressure, but in 19 per cent blood pressure went up. Pulse rate increased in 25 per cent of patients. This indicates that the placebo device is only useful if the two groups could be equated for suggestibility. At present I know of no way in which this can be done.

For this reason 9 points were disregarded, being lopped off from the lower end of the distribution table of differences. It was hoped that this would help to eliminate to a considerable degree minor variances in behavior due to suggestibility.

Figure 8 shows graphically the changes in ward-behavior ratings that took place over a period of 18 weeks. There were two ratings given before medication, and their average level is marked on the scale to the left. For purposes of readability, the results are shown in two graphs placed side by side. One half of the figure includes mental confusion, untidiness (in the hospital sense), hallucinations, and delusions. The parallel graph shows disorders of speech, negativism, aggressiveness, restlessness or physical overactivity, and emotional level (mainly depression). Had these all been shown in the one graph, lines would have too closely overlapped. If the graphs are divided transversely, then the top half of each would contain the behavior traits which are thought by psychiatric aides to be most noticeable in the reactions of chronic psychotic patients, since these receive the highest or worst ratings.

The steepness of descent of the lines indicates the traits that are

changed most radically by chlorpromazine. They are seen to be hallucinations, untidiness, negativism (mainly in ward cooperation), aggressiveness, physical overactivity, and delusions. The greatest

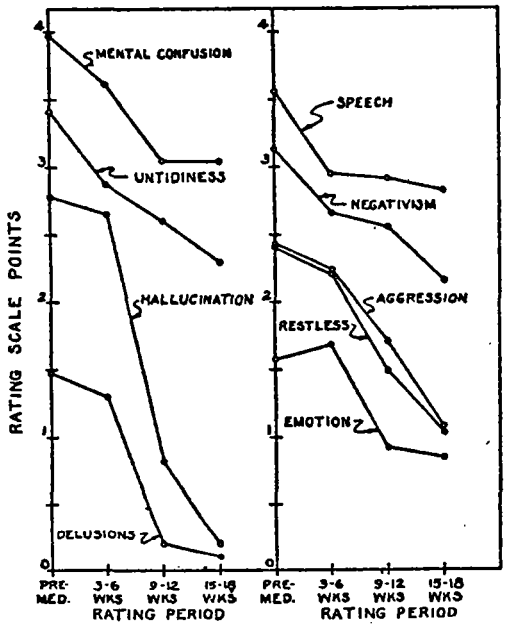


FIG. 8. Changes in Behavior Ratings, after Chlorpromazine

change of all occurs in hallucinations, but it will be noted that during the first six weeks of drug administration the changes were relatively slow. As a matter of fact, mental depression actually increased during this period. The similarity of pattern of decline of hallucinations and delusional behavior to aggressiveness and restlessness suggests an association between these types of reactions. This may be expressed otherwise by saying that as the patients regained contact with reality their social behavior improved. They were no longer as intolerant of others' viewpoints and found themselves more in tune with their fellows. Such changes make patients of this type easier to live with and are of importance in relation to a type of social adjustment not represented in Maze Test performance. Becoming less sensitive to environment and diminished alertness constitute negative traits. Improved Maze scores reflect more dynamic or positive adaptations. Lifting the load of anxiety makes a patient

saner but it may not be mirrored in mental age scores. If anywhere, the change would be more likely to be reflected in qualitative execution of the tests. Figure 9 shows graphically individual 6-week

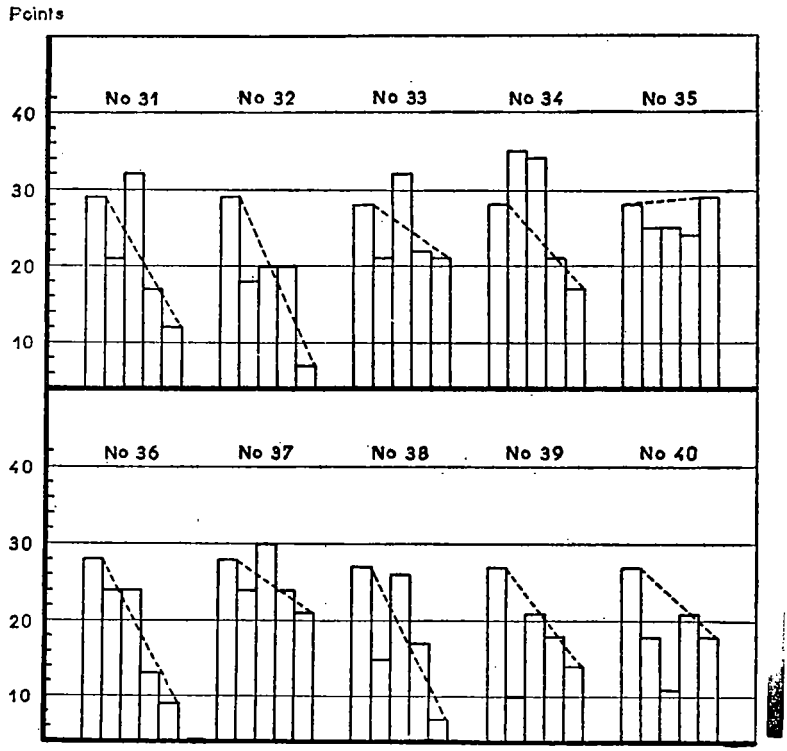


FIG. 9. Individual Rating Changes after Chlorpromazine. At Six Week Intervals

(Slope of dotted lines indicates amount of behavior change from premedication ratings to final postmedication ratings at 18 weeks.)

changes, the downward slope of dotted lines indicating degree of improvement. The fluctuations are quite obvious.

This discussion of the effects of chlorpromazine on social behavior has special relevancy to the Maze Test because of a surprising and partly accidental discovery that deficits in performance followed its use. As part of our research design for this psychopharmacological study, we tried to apply the original Maze before medication. However, we found that the test was applicable only to 13 out of the 50 chronically psychotic subjects to whom the drug was to be

administered. Though the test is usually accepted by hospital patients, factors of age, deterioration, physical disabilities, plus occasional unresponsiveness, combined to make psychometry impracticable for the rest of the group. In some of the patients tested, scores were near the floor of the test. Hence very little emphasis was at first placed on this phase of the investigation.

When, however, the drug seemed to bring about startling changes in social adjustment, at least within a hospital environment, it was decided to re-apply the test, using the Extension or practice-controlled form for the postmedication examination. This, it should be said, was done with full confidence that the improved sociability would be reflected in improvement in test scores.

It is true that a parallel between psychosurgery and tranquilizing drug therapy had been suggested by research workers who had studied the effects of the latter. In a review, Dundee [5] made the following statement:

> The mental effects of prolonged administration of chlorpromazine have been called "pharmacological frontal lobotomy" (Terrain, 1952). Patients show a lack of spontaneous interest in their surroundings, are generally immobile, and at first glance seem to be heavily drugged. However, the higher psychic functions are preserved to a remarkable degree and patients are capable of sustained attention and concentration.

Other similar references are to be found, and seem to be based on some rather general effects. Lobotomy and chlorpromazine both are used to relieve intractable pain; both diminish anxiety or exaggerated self-concern; both frequently are followed by improved appetite and marked increase in body weight. But these changes can only be rather vaguely summarized or described, as are also similar changes in affect, notably a dulled emotional response to delusional or hallucinatory ideas that were formerly reacted to much more intensely or overtly.

It is noteworthy that Landis and Erlick,[6] who carried out a study of Maze reactions in Columbia-Greystone subjects, noticed some general characteristics in patients' attitudes while taking the test. They say, "It is as though there were a decrease in the vigilance or wakefulness of the patient, the changes being as varied as those which might occur in any individual that was drowsy or sleepy while taking a test." * It is true that drowsiness is also initially one of the effects of chlorpromazine, while loss of initiative, especially as regards aggressiveness, is one of the most noticeable changes in personality, but except for these, no other deficits were

* This effect suggests that the reticular formation is involved.

easily apparent to ordinary clinical observation. It was not until the Maze was applied after medication that the similarity of its effects to lobotomy and other forms of psychosurgery became more pertinent.

Nine additional cases who had been tested prior to medication were found in other wards, raising the total in the Maze experimental group to 22. This number may seem small, but it was difficult to find new patients who had not had tranquilizers before entering the hospital. Others on admission were acute cases and had been put on the drugs by ward physicians immediately to calm them down. These nine cases could not be all considered chronic psychotics, and the medication had been given for 6 weeks instead of 18 weeks as was the case with our chronic patients.

The results for these 22 cases were surprising. There was a net loss in Maze scores (algebraic sum of gains and losses divided by N) of 2.06 years in test age. The percentage declining in Maze scores was 68, 14 per cent remained the same, and 4 patients, or 18 per cent, actually gained in Maze performance. In the case of two out of four who improved their Maze scores, there was considerable doubt as to whether they had taken the drug regularly. They admitted that when the aide was not looking, they sometimes disposed of the pills down the toilet. It was found that strict supervision of patients was easier in the closed wards than in the admission unit. Results are given in Table 19, which also gives comparative figures for lobotomy cases.

I have offered some observations of the test reactions of patients to the Maze after medication. There was sometimes observable what has been called "shatter effect," e.g., good planning and a careful approach up to about the eleven year Maze design, followed thereafter by a rather sudden decrease in vigilance resulting in failure. This deficit did not seem due to weakened motivation. Patients sometimes apologized for their mistakes, but on trivial grounds. A couple of them said they were sorry for wasting so much of the examiner's paper. As far as I could judge, subjects actually put forth their best efforts, but any concern for failure was only of superficial depth. This, too, is very typical of lobotomy patients. It reminded me of Jacobsen's lobotomized monkeys, who, when they failed to secure food in a test situation, did not become angry or disturbed as before the operation, but accepted frustration with apathy or indifference. Chlorpromazine patients also seemed to regard frustrating experiences without direct self-reference. One could not say that they were indifferent to failures in the Maze. They tried to do well but if they did not succeed, it was not worth worry-

ing about. The reason for failure seemed to be a decided impairment in planfulness. Of those who declined in score (15 cases), eleven lost 2 or more years of test age, and of these six lost five years or more. This is a very marked decline in planfulness.

TABLE 19

Maze Scores Before and After Lobotomy and Chlorpromazine

Case No.	Lob. Porteus—Kepner			Lob. Porteus—Peters			Chlorpromazine		
	Pre-Op.	Post-Op.	Diff.	Pre-Op.	Post-Op.	Diff.	Pre-Med.	Post-Med.	Diff.
1	13.5	14.5	+1.0	8.5	10.5	+2.0	6.0	10.5	+4.5
2	14.5	15.5	+0.5	10.5	14.5	+1.0	6.0	6.5	+0.5
3	8.5	8.5	0.	14.5	15.0	+0.5	14.0	14.5	+0.5
4	6.0	6.0	0.	11.0	11.0	0.	16.0	16.5	+0.5
5	14.5	13.5	−1.0	10.5	10.5	0.	16.0	16.0	0.
6	6.0	5.0	−1.0	10.0	9.5	−0.5	6.0	6.0	0.
7	12.5	11.5	−1.0	7.0	6.5	−0.5	16.5	16.5	0.
8	5.0	4.0	−1.0	10.0	9.0	−1.0	6.5	6.0	−0.5
9	11.5	9.5	−2.0	8.5	7.5	−1.0	13.5	13.0	−0.5
10	13.0	11.0	−2.0	12.0	9.0	−3.0	8.0	7.5	−0.5
11	9.0	7.0	−2.0	9.5	6.0	−3.5	12.0	9.5	−2.5
12	14.0	11.0	−3.0	11.0	7.0	−4.0	15.5	12.5	−3.0
13	14.0	10.5	−3.5	14.0	9.5	−4.5	15.0	11.5	−3.5
14	11.0	7.0	−4.0	14.0	9.5	−4.5	15.0	10.0	−5.0
15	13.0	8.5	−4.5	13.0	8.5	−4.5	14.0	7.5	−6.5
16	14.0	9.5	−4.5	13.5	8.0	−5.5	9.0*	7.5	−1.5
17	13.5	8.5	−5.0				9.0*	7.0	−2.0
18							11.0*	8.5	−2.5
19							13.5*	8.5	−5.0
20							15.5*	10.0	−5.5
21							15.0*	9.0	−6.0
22							15.5*	8.5	−7.0
	Loss 1.91 Years 76.5% Cases			Loss 1.81 Years 68.75% Cases			Loss 2.06 Years 68.2% Cases		

* Female patients.

In reporting this study, I have commented upon the fact that the impairment occurred at all levels in both psychosurgical and chlorpromazine cases, with the greatest decline naturally being apparent when the premedication or preoperative score was high. Furthermore, the amounts of deficit varied greatly in different individuals. It seems certain, therefore, that the brain mechanisms underlying Maze performance are extremely complex with probably several areas involved in its mediation and that the distribution of the

effects of either surgical or drug therapy are very uneven. The severe deficits recorded in some cases certainly resembled those found in superior cortical ablations, posterior venous ligation, and thalamotomy.

Though difficult to place in definite categories, the resemblances between chlorpromazine effects and psychosurgical effects are sensed by the experienced examiner. Some observations were summed up in a recent publication: [7]

> However, there is no typical or general reaction to the Maze by either psychosurgical or chlorpromazine patients but a great variety of attitudes. Some are nonchalant in the face of failure, some are cautious, conscientious workers. Some appear to think a great deal depends upon their success or failure, but most seem to enjoy the experience; some are apologetic (for mistakes), others are emotionally flat.

One tendency that can be measured is perseveration, as far as it is exemplified by entering the same blind alleys in repeated trials. In the Porteus-Peters study, 21 per cent of lobotomy patients had two or more repetitive errors in their preoperative examination, 37 per cent in their postoperative tests. Chlorpromazine patients examined prior to medication also had 20 per cent repetitive errors, 45 per cent after medication. The whole group averaged 1.1 repeated errors per individual before, 1.55 after. Thus, while psychotic patients tend to make more repetitions than normals, the trend is accentuated after either lobotomy or chlorpromazine, particularly in the latter group of patients. It will be remembered that Petrie [8] also found a greater tendency toward "double mistakes" after frontal leucotomy.

It should be mentioned that the percentages of individuals showing decline in Maze scores are almost identical for both therapies—in the chlorpromazine group 68 per cent, in the second postoperated lobotomy group 69 per cent. However, some psychosurgical patients obtained lower scores in a subsequent testing so that the percentage of those whose Maze level declined at some postoperative testing finally reached 82 per cent.

The next step was to increase the size of the experimental group and to compare their scores with those of a control group. The difficulties of setting up adequate controls have already been discussed. If, as seemed certain by ward behavior observations, anxiety, antagonism, restlessness, and suspicion had decreased in the treated group and had not changed in the control group, equating for such temperamental effects could not be adequately done. The changes noted should have the effect of improving Maze test reactions.

Hence, any decline in the experimental group would be the more remarkable. They could be interpreted as losses in initiative and planning capacity in excess of any favorable temperamental gains, due to the tranquilizing effects. The continued studies were reported by Porteus and Barclay.[9]

The number of cases in the experimental group was later raised to 35, and 25 other patients were used as controls, but with due recognition of the device's weakness. The original Maze was used before medication, the Extension Maze after medication. The results are given in Table 20. It will be seen that the means and S.D.s of the Standard or original Maze were almost the same for both groups prior to medication.

But the experimental group after medication lost 1.89 years, whereas the control group changed 0.1 of a year. Evidently practice was controlled, judging by this latter near-equivalence.

We then decided to match twenty pairs of both groups with identical standard Maze scores, rather than to rely on mean equivalence of drug cases and controls. The results are also shown in Table 20. Under this procedure the experimental group lost 2.2 years of test age, while the control group gained 0.2 years.

The Maze Extension was then repeated a second time in reversed position, 21 patients being then available for testing and comparison. Whereas the deficit for the experimental group then changed from 1.5 years to 2.09 years, the control group changed from a loss of .04 of a year on the first Extension testing to a gain of 1.07 years on the inverted Extension. The total difference between the two groups finally was 13.09 years minus 9.91 years, or 3.18 years of test age. The figures are given in Table 21. A third postmedication Maze application with the Extension series rotated 90° showed practice effects, the control group reaching nearly to the ceiling of the test. The experimental group, however, had mean scores still well below their premedication level.

As to what areas in the brain are affected by chlorpromazine, there is little experimental evidence as yet available. In 1952, Terzain concluded that it depresses the action of the reticular formation, which seems to be related to mental alertness. This may be so, which would account in part for the Maze deficits. It has been noted that the chlorpromazine patient wakes up after the drug without any continuing somnolence or mental confusion, a quite different reaction to the individual awakened after barbiturates. Immediate alertness is said to follow reticular stimulation.

From analyzing the effects of chlorpromazine on the central and autonomic systems, Aron in 1954 concluded that the site of its action

is probably in the hypothalamus. However, the greater Maze impairment that follows ablation of the cortex of areas 8 and 9 would argue that superior cortical regions are involved in Maze perform-

TABLE 20

TEST DEFICITS AFTER CHLORPROMAZINE
AGE CHANGES IN STANDARD AND EXTENSION MAZES

Maze Scores of Experimental and Control Groups			Matched Experimental and Control Cases		
Measure	Experimental	Control	Measure	Experimental	Control
N	35	25	*N*	20	20
Standard Maze			Standard Maze		
Mean	11.90	11.96	Mean	11.85	11.85
SD	3.42	3.34	*SD*	3.50	3.50
Extension Maze			Extension Maze		
Mean	10.01	11.86	Mean	9.65	12.05
SD	3.58	3.42	*SD*	3.20	3.60
Difference	−1.89	−0.10	Difference	−2.20	+0.20
t	2.26*		*t*	2.10*	.20
F	4.99*				
z[a]	2.21*				

* Significant at .05 level.
[a] Mann-Whitney test.

ance. In my opinion, the facts merely indicate that successful Maze solutions are mediated by a very complex organization rather than any limited areas, but the abilities that underlie maze-threading are so related to self-preservative reactions that the most primitive, yet essential, centroid neurobiotaxical brain centers are involved. This idea has been discussed earlier.[10]

Admittedly, these matters are speculative rather than factual considerations, but the chlorpromazine results may be accepted as an important link in the chain of evidence supporting the claim that the Maze should be regarded as having, more than any other test, a psychosomatic reference. The losses that follow lobotomy, the more serious the more posterior the cut; the similar deficits that result from topectomy, the more marked if the ablations are superior and posterior rather than inferior and anterior, the former losses being apparently permanent; the drastic declines that follow temporal

lobotomy; the loss in effective performance in the case of severe Vitamin-B deficiency and the Maze impairment associated with prolonged chlorpromazine medication together form a strong chain of presumptive evidence. Furthermore, when other test scales are used following such approaches, they are either completely insensitive or only slightly or inconsistently sensitive. The natural implication of such a chain of evidence unless it is blindly ignored, is to admit the Maze to equal significance with these other scales, in all cases where mental diagnosis is effected by psychometry.

Tables 20 and 21 give the results of a comparison of experimental and control groups. Leaving the latter out of consideration it will be seen that the Maze deficiency for 35 cases after chlorpromazine amounted to 1.89 years. This is greater than the amount of defect for lobotomy cases but it must be remembered that in that study the original Maze was repeated whereas in the more recent work the practice-free Extension was used as the postmedication measure. For the 20 matched cases the deficiency was 2.2 years in postmedication scores. Table 21 indicated a further drop in Maze performance

TABLE 21

MAZE SCORES—REPEATED APPLICATIONS

Measure	Experimental	Control
N	21	21
Standard Maze		
Mean	12.00	12.02
SD	3.60	3.24
Extension Maze 1		
Mean	10.50	11.98
SD	3.89	3.29
Difference	−1.50	−0.04
Extension Maze 1 (Inverted)		
Mean	9.91	13.09
SD	3.61	2.95
Difference	−2.09	1.07
t	1.92*	1.17

* Significant at .05 level.

to 9.91 years when the Extension Maze was inverted whereas the control group had risen to 13.09 years.

The fact of Maze impairment after the drug seems well substantiated.

CHAPTER REFERENCES

[1] Waite, J. H. and Neilson, I. L., "A Study of the Effects of Hookworm Infection upon the Mental Development of North Queensland School Children," *Med. J. Australia.* No. 1, pp. 1–18, January 4, 1919.

[2] Porteus, S. D., "Specific Behavior Changes Following Chlorpromazine," *J. Consult. Psychol.*, **21**, 257–63, 1957.

[3] Porteus, S. D., "Variations Specifiques du Comportement sous l'Effect de la Chlorpromazine," *Revue de Psychol. Applique* **6**, 187–202, 1956.

[4] Rosenthal, D. and Frank, J. D., "Psychotherapy and the Placebo Effect," *Psychol. Bull.*, **53**, 294–302, 1956.

[5] Dundee, J. W., "A Review of Chlorpromazine Hydrochloride," *Brit. J. Anaesthesia,* **26**, 358–80, 1954.

[6] Landis, C. and Erlick D., "An Analysis of the Porteus Maze Test as Affected by Psychosurgery," *Amer. J. Psychol.*, **63**, 557–566, 1950.

[7] Porteus, S. D., "Maze Test Reactions after Chlorpromazine," *J. Consult. Psychol.*, **21**, 15–21, 1957.

[8] Petrie, A., *loc. cit.*

[9] Porteus, S. D. and Barclay, John E., "A Further Note on Chlorpromazine: Maze Reactions," *J. Consult. Psychol.*, **21**, 297–99, 1957.

[10] Porteus, S. D., "Some Common-Sense Implications of Psychosurgery," *Brit. J. Med. Psychol.*, **28**, 167–176, 1955.

CHAPTER XI

History, Application and Scoring

Historical Summary

For forty-four years the Porteus Maze Test has continued in clinical use without major changes in character and form. Yet in the past ten years most important new uses have been found for this measure and each has added greatly to the body of evidence for its validity. It is surely unique that a psychological tool devised almost half a century ago should be in increasing demand today. Because such a long period of time is involved, a brief historical summary should be provided for the information of new users of the Maze.

General intelligence tests, such as its leading contemporary the Binet-Simon Scale, have either fallen into disuse or have been changed so much in repeated revisions as to be unrecognizable as far as resemblance to the original scale is concerned. Of late years Binet tests have suffered a decline in interest and presumed value. Neither of the above considerations apply to the Maze.

In clinical usage, the Binet tends to be replaced by the Wechsler-Bellevue Scale. However, its selection of subtests and the principles upon which all "general intelligence" measures are founded have recently come under such weighty and severe criticism that drastic revision of the scale seems called for. Undoubtedly, clinical psychologists need an easily applied measure of children's generalized ability, but in my opinion an adult scale, arranged on a similar hodge-podge principle, must suffer an inevitable decline in estimation and use.

On the other hand, the Maze escapes censure on at least one essential point. It does not aim to be a general measure, but is directed toward the thorough exploration of a segment of mental behavior which seems to be best described as "planfulness." It is recognized at the outset that this is a trait-complex in which intellectual, emotional and/or temperamental factors are involved. It is

directed towards measuring not high level planning but foresight (and the tendency to use it) in matters of every-day experience. The mental activity which appears to be most directly involved in success in the Porteus Maze would seem to be vigilance or mental alertness, especially as concerns anticipatory reactions. For the present at least, it does not compete with any other test.

My main thesis in this book is that foresight, planning, anticipation, prehearsal, by whatever name it is called, is an integral part of any intelligent act that has not yet passed the learning stage and become automatic. Because of the essential relation of planning capacity to intelligence, its specific evaluation is necessary to psychometrics as a branch of clinical psychology.

Obviously, the usefulness of the Maze for this purpose depends on its proven validity. I have repeatedly expressed the hope that more research will be directed towards the development of better measures of planfulness.

The changes that have occurred in the Porteus Maze during its long history are mainly concerned with details of scoring. Changes in form that have taken place in the Standard Maze series as first published in 1915 are not considerable. Except for the V, VI and VII year tests, the designs are very similiar to those in use today. The scoring range originally extended from three to thirteen years of test age.

In 1919, what is still known as the Vineland Revision [1] was published along with general instructions for application and scoring. Credit was allotted up to 14 years of age and a system of allowing a certain number of repeated trials was adopted.

In 1924, a booklet entitled *Guide to the Porteus Maze Test* [2] appeared. There was, as is the case in all Maze series since 1915, no test design for 13 years, but credits were added for successful trials in tests XII and XIV, which together with those for two adult tests enabled S to earn a total score of 18 years. The use of the adult tests was optional.

In 1933, a book on the use of the Maze Test [3] was published. The old design for Year VI was changed and a single adult test was included in the series. The maximum score allotted if all tests up to and including Year XIV were passed on the first trial was 15 years; if the adult test was similarly passed, 17 years. As in the earlier series, the absence of any test design for 13 years was compensated for by allowing 2 years' credit for first trial successes in Years XII and XIV.

In 1950, the modern practice of administering and scoring the Maze, with only minor changes from the 1933 edition, was set forth

in another volume.[4] These changes concerned mainly the qualitative scoring introduced in 1942. In 1955, a booklet[5] described a new practice-controlled series known as the Maze Extension, the application of which will be described later.

It is important to note that instructions and scoring rules as given in this present volume supersede all former directions.

Rules for General Application

In summary the present Porteus Maze Test consists of two series. The first is used with children aged three to twelve years of age. It contains 11 test designs, ending with Year XIV. Since 2 years' scoring credit can again be obtained in each of the XII and XIV year tests if passed on the first trial, the ceiling score is 15 years.

The tests for Years III and IV are not really mazes, the problem being to draw between the double guide lines with a minimum of line crossings. However, the ability to anticipate changes in direction is involved. Except for very young or very dull children, testing usually begins with Year V. In cases where it is obvious that S does not comprehend the instructions, the design for Year V may be used for demonstration or fore-practice purposes.

Adolescent children 13 years and above and adults are given the same series of tests with the addition of an adult test, designated Adult I. The ceiling score, if all tests are passed on the first trial, is 17 years. For normal subjects, testing begins with Year VII, but if through language disabilities or for other reasons S fails to comprehend the instructions, a test lower than Year VII may be used for demonstration.

Two trials are allowed, if needed, on every test up to and including Year XI; four trials in Years XII, XIV and Adult I.

An *unsuccessful trial* is recorded as soon as S enters to any noticeable distance or degree into a blind alley. Entrance is defined as the crossing of an imaginary line across the opening leading into the blocked path or street. On no account should S be allowed to retrace his course or continue with the same design after entering a blind alley.

The only continued drawing allowed is along the path S has already entered. The design is taken from him and a new trial begun when he stops or otherwise indicates that he is blocked.

It should be noted that an unsuccessful trial does not constitute *failure.* If S does not succeed in the allotted number of trials—two for each test up to Year XI, four for any higher test—failure is recorded for that particular test.

No additional credit is allotted if failures (as defined in the pre-

ceding paragraph) occur in any three years in the whole series or in two successive years above Year IX. For mental age reckoning, testing normally ceases after such failures, but if a complete qualitative record is desired, S may be allowed to work through the additional higher tests but no credit is allowed for success in any of these.

To diminish the likelihood of accidental successes, the rule for inversion operates. If after recorded failure in any test S passes the next higher test in the allotted number of trials, this design is inverted and the test repeated as if it was a new application. For example, if the IX year test is failed in two trials and the X year is passed, this latter must be inverted and given again. In scoring the test age, the worse of the two records, either the original or inverted, is used.

Note: Neglecting to invert a test successfully passed when failure has occurred in the test immediately below it in the series is one of the commonest administrative mistakes made by the inexperienced examiner. The other most important error is to allow S to continue tracing through a maze design after entering a blind alley. "Wrong choice" mistakes cannot be self-corrected. Even if it is the last possible error, the test design must be withdrawn and a new trial begun. The examiner, otherwise, has no means of knowing whether a wrong choice may not be made at the same or another point in the design.

A scoring rule applying to all Maze series, whether Standard (original), Extension or a third form now being standardized, is to take the highest test recorded as passed under the previously set forth conditions, and then deduct a half year for each unsuccessful trial throughout the whole series.

Procedures for application of the test series for adolescents and adults remain the same as for children. The examiner may encourage S if necessary but must never help him.

Except for tests up to and including Year VII, the examiner does not point out the exit. In some of these higher tests S frequently asks, "Where do I come out?" He should then be told that he must try and find his own way out.

Occasionally in one of the more complex designs S will sometimes stop and announce that he is blocked or that there is no way out. After a delay of a couple of seconds, the examiner should remove the test design and record an unsuccessful trial. Sometimes this occurs when S is actually on the right track. If he has not used up all his allotted trials, a new test blank of the same design is presented with the comment that there *is* a way out.

S is allowed all the time he needs for preconsideration of any maze, but he must be forbidden to trace a course in the air with finger or pencil. Also, to prevent undue practice, only the allotted number of trials should be given.

Crossing lines except when S cuts across to reach an opening instead of going around is a qualitative not a test age error. Lifting the pencil from the paper contrary to initial instructions is also a qualitative error. Occasionally a child forgets and lifts the pencil and assumes that by so doing he has failed the test. He should be informed otherwise and allowed to continue.

For additional emphasis the testing rules have been summarized in concise form.

Summary of Testing Rules

1. Begin testing of all children five to twelve years of age with Year V and end with Year XIV. Cases, thirteen years and over (including adults), begin with Year VII and end with Adult I.
2. Continue recorded testing until S has failed in any three tests or in two successive tests in Years X or above.
3. Two trials are given in each test up to Year XI, four trials in Years XII, XIV and Adult; a test *failure* is recorded if S does not find his way through a maze in the allotted number of trials. Each blind alley entered is an *unsuccessful trial.*
4. S must never be allowed to correct a "wrong choice" error and continue tracing through the same design.
5. Tracing the course in the air or proceeding with more than the allotted number of trials is forbidden.
6. Crossing a line, except in deliberately cutting across to reach an opening, is not a test age error.
7. If S succeeds in a test after failure has been recorded in the test immediately below it, the test design is inverted and repeated. The worse record is used in totaling deductions due to unsuccessful trials.

Detailed Instructions

The examiner should sit opposite S and hold the Maze in position with the tips of the fingers resting on the top of the test blank. A moderately soft pencil of medium bluntness should be provided for tracing the maze. A sheet of paper should be placed under the test blank so that irregularities in the surface of the testing table will not affect performance.

Year III

Examiner says:

"Look at these lines on the paper. I want you to take this pencil and draw around between the lines just as carefully as you can without touching them. Be sure to keep the pencil between the lines like this." (Examiner demonstrates by drawing about an inch along the path, beginning at the point marked "S," and proceeding in the direction of the arrow. If the child does not understand, examiner may guide the child's hand and pencil down to the first angle in the design.)

Two trials are allowed and three year credit is given if on either trial there are not more than 3 places where the guide lines are crossed.

Year IV

Examiner says:

"Do this one just the same way. Begin here and draw right around between the lines but try not to cross any lines." (Examiner indicates the starting point and the direction in which the pencil is to move.)

Two trials are allowed and four year credit is given if on either trial there are not more than two errors made by crossing lines.

Year V

Examiner says:

"This is what is called a maze and you must try to find your way through it. You can imagine that these lines are all stone walls and that a rat came along and saw this hole." (Examiner points to entrance to the maze.)

"Then he found his way between the walls until he came to the cheese. Now I want you to draw a line to show me where the rat went to get the cheese. But you must be very careful not to run into any lines or go into any place that is blocked at the end. The rat cannot turn around and come back."

"*And one thing more you must remember*—you can stop anywhere and look as long as you want to, but try not to lift your pencil until you have found your way out. Now take this pencil and begin here (Examiner points with the finger to the initial arrow) and draw for me the way the rat went to reach the cheese here." (Examiner points to exit.)

If S is deaf or does not understand English, the examiner should point to the rat and the cheese, then take the pencil and draw through the maze, hesitating at each choice point before going on. S is then given a blank test and encouraged to do the same thing.

For primitive individuals, both the tests for Years V and VI may be used as fore-practice.

Two trials are allowed.

Year VI

Examiner says:

"You are to begin here (Examiner points with the finger to the end of the arrow) and show me how to find your way out here." (Examiner points to exit.)

Two trials allowed.

Year VII (Children's series)

Examiner says:

"Begin here and find your way out here." (Examiner points to entrance and exit.)

Two trials allowed.

Year VII (For individuals aged 13 years and over.)

As this is the starting point in the testing of older Ss, the instructions for this test must be given in full detail for these older cases.

Examiner says:

"This is what is called a maze. I want you to suppose that these are streets and all the lines are stone walls and you have to try and find your way out. So you take this pencil and start here (Examiner points with his finger, not the pencil, to the starting arrow) and find your way through the maze to here." (Examiner points to place of exit.)

"But you must be very careful not to go the slightest distance into a blind alley because you are not allowed to turn around or come out once you enter a blocked street. If you do, you must stop and another maze blank will be given you."

"This is not a speed test. You can stop and look anywhere as long as you like but try not to lift the pencil off the paper until you are right outside the maze, and do not bump into any walls. You may start any time now and draw through the maze."

Note: The place of exit is not pointed out in any higher test.

Two trials allowed.

Years VIII, IX, X

Examiner says:

"Begin here and find your way out." (Examiner points to the starting arrow but not to the exit.) He replies to questions as to the place of exit by telling S he must find his own way through the maze.

Two trials allowed in each test.

Years XI, XII, XIV and Adult I

Examiner says:

"Begin here in the middle and find your way out." (Examiner points to the letter S in the center of the test, but does not indicate point of exit.)

Two trials allowed in Year XI. Four trials allowed in Years XII, XIV and Adult.

Scoring Procedures and Examples

The system of scoring has been somewhat simplified. It consists in adding up all the unsuccessful trials taken by S in working through the series and deducting a half year of test age for each such trial from the score allotted to the highest test passed. The rule with regard to not allowing test age credits for any tests given after three complete failures in the series or two successive failures above Year IX must of course be observed.

The following samples of scored records are given by way of illustration.

Example 1 (Child under 13 years—one inversion)

Test	Record	Deductions
V	2 trials	½ year
VI	1 trial	
VII	2 trials	½ year
VIII	Fail	1 year
IX	1 trial	
IX inverted	2 trials	½ year
X	1 trial	
XI	1 "	
XII	3 trials	1 year
XIV	4 "	1½ years
	Total deductions	5 years

Highest credit is 15 years, less 5 years for unsuccessful trials. Test age = 10 years. Year IX was inverted because of failure in VIII, the test immediately below in the series.

Example 2 (Child 12 years of age)

Test	Record	Deductions
V	1 trial	
VI	1 "	
VII	1 "	
VIII	2 trials	½ year
IX	2 "	½ year
X	2 "	½ year
XI	Fail	
XII	4 trials (no credit)	
XII inverted	Fail	
XIV	2 trials (no credit)	
	Total deductions	1½ years

Highest credit 10 years, minus 1½ years for unsuccessful trials. Test age 8½ years. Year XII was passed on the fourth trial but because Year XI had been failed, XII was inverted and failed. This made two successive failures above the IX, hence no additional credits could be given. Test XIV was given for qualitative scoring only and did not count in the test age.

Example 3 (Child under 13 years)

Test	Record	Deductions
V	1 trial	
VI	1 "	
VII	2 trials	½ year
VIII	2 "	½ year
IX	1 trial	
X	Fail	1 year
XI	2 trials	½ year
XI inverted	1 trial	
XII	4 trials	1½ years
XIV	Fail	2 years
	Total deductions	6 years

Highest credit 15 years, minus 6 years, test age 9 years. Test XI was inverted after failure in Test X. Since S qualified for application of Year XIV, the ceiling score was 15 years.

Example 4 (Case 13 years or over)

Test	Record	Deductions
VII	1 trial	
VIII	2 trials	½ year
IX	Fail	1 year
X	1 trial	
X inverted	Fail	1 year
XI	2 trials	½ year
XI inverted	1 trial	
XII	3 trials	1 year
XIV	3 "	1 year
Adult I	2 "	½ year
	Total deductions	5½ years

Highest credit 17 years, minus 5½ years, test age 11½ years. Test X was inverted because of failure in IX. Test XI was also inverted after success because of failure in X, the next lower test. Since the two successive failures did not both occur in Test X or above (Rule 2), application continued up to the Adult level.

Example 5 (Case 16 years of age)

Test	Record	Deductions
VII	1 trial	
VIII	Fail	1 year
IX	Fail	1 "
X	1 trial	
X inverted	2 trials	½ year
XI	2 "	½ year
XII	3 "	1 year
XIV	Fail	2 years
Adult I	4 trials (no credit)	
	Total deductions	6 years

Highest credit 15 years, minus 6 years, test age 9 years. Since failure in XIV was the third in the series, no credit could be given to Adult I which was given for qualitative scoring purposes only. The ceiling score was 15, since S qualified for application of Test XIV.

Qualitative Scoring

The qualitative as distinct from the mental age scoring is concerned with errors in drawing or execution rather than in planning. In other words, it has to do with S's ability to follow instructions with regard to crossing lines and lifting the pencil, and also takes into account the general neatness of the performance and whether quantitative or "wrong choice" mistakes take place near the beginning or end of the test design. The score is intended to reveal any haphazard, impulsive, or over-confident habits of action, or a tendency to become so absorbed in the task of finding the way through the maze as to neglect other considerations. The weighting of errors is based on the comparative performance of delinquents, for whom the qualitative score is particularly revealing.

No criticism of the performance is given by the examiner and every error in execution must be rigorously scored. Leniency may change the picture considerably. The higher the Q-score, the worse the record.

Details of Scoring

The following errors are marked on the face sheet, weighted as indicated and totalled for the qualitative score.

1. Any wrong choice or test age error occurring in the first third

of the design, i.e. at or before the arrow marked on each sample maze. Weighting 2.

2. Any wrong choice error occurring in the last third of the design, i.e. at or after the arrow marked on each sample maze. Weighting 1.

3. Each cut corner. The error must occur while actually turning a corner. Weighting 1. A line crossed near a corner but not related to turning is scored as a crossed line.

4. Wherever the pencil mark *touches* a printed line in a maze, other than in turning a corner, it is scored as a crossed line. Touching a line when going through an opening is similarly scored. Weighting 2.

5. Lifting the pencil except at the exit is a qualitative error. The warning against lifting the pencil is given in the initial instructions for Test V (children 12 years of age or under) and in the instructions for Test VII with older cases. The warning should be repeated again after 5 liftings in a single test, or after a total of 10 liftings in any combination of tests. Weighting 3.

6. "Wavy lines" are recorded against S if his general performance is as irregular in appearance or worse than the performance given as a scoring sample of the test in question.

7. "Changed direction" (formerly wrong direction) is the name given to an error where it is evident that S started to go into a blocked path but stopped the pencil before crossing an imaginary line across a "wrong choice" opening. If such a line is crossed, it would of course be scored as an unsuccessful trial. Weighting 1.

8. For children under 13 years, any qualitative error occurring in Test VI obtains an added penalty of one point for each such error. For cases 13 years and over, this added penalty applies to Test VII. A place in the face sheet provides space for recording. Weighting 1.

The Q-score is the sum of the weighted scores in the above list, but there are two important restrictions. *The maximum weighted error score recorded against any individual in any single item is 48.* This means that only 24 crossed lines or 16 lift pencils are recorded. It was found that in some cases, particularly delinquents, so many errors occurred that the score lost meaning for comparative purposes. There are no real distinctions between an individual scoring 120 and one scoring 150. For that reason the maximum Q-score is 100.

Obviously, S should not be made aware that his performance is being scored. However, the number of lift pencils in any design should be counted and the notation LP followed by the number should be marked on the test design.

Changes in Former Procedures

For those who are already familiar with the use of the Porteus Maze, the main changes in procedures may be briefly mentioned.

The instruction given to examiners to draw carefully around the first turn in the first maze presented (Test V for cases under 13 years; Test VII for older cases) has been dropped. This is because the type and place of beginning is very important in observing consistency of performance and in reckoning the similarities score. For the same reason, the examiner points at starting places and exits with the finger. Using the pencil would mark the place too exactly.

For similar reasons, instructions are no longer given in terms of driving an automobile through streets. It is desirable for the examiner to note the occurrence of right angle turns and it may occur to S that a car cannot make a right angle turn. The principle in mind is to interfere as little as possible with S's ordinary style of response.

Formerly, additional penalties were imposed in qualitative scoring after a warning against lifting the pencil had been repeated. These no longer apply.

In test age scoring, the rule regarding successive failures in two year tests now applies above Year IX instead of Year VIII.

Instead of adding credits for successes in tests, a uniform, more convenient system of deductions for unsuccessful trials is used.

Finally, the table for reckoning test quotients in the Maze allows cases to obtain a maximum T.Q. of 135 at each age level. In my opinion, T.Q.s below 30 and above 135 are of little comparative significance. The practice of including months of chronological age in divisors to obtain ratios between mental ages and chronological ages also seems unnecessary, as it confers an air of exactitude which is not justifiable. In the tables of T.Q.s appearing in the present volume, chronological ages are given in half years up to and including 10 years, in twelve-month intervals thereafter.

Though the previous tabulation of test quotients is no longer recommended for use, the old tables are included for the benefit of those psychologists who have long term research projects in operation.

Maze Test Extension

The Maze Test Extension series consists of eight test designs standardized in 1955 so as to give scores equivalent to those of the original or standard series. This Extension was devised in order to control practice effects whenever it is necessary to repeat the

Maze after an initial application. As an example of this use, we may instance studies of patients before and after some special therapeutic measures, such as psychosurgery or administration of tranquilizing drugs, where a practice-free form of the Maze is required. A second practice-free series has also now been standardized and will be published in due course. Testing in the Extension series begins at Year VII and the instructions for application, scoring and T.Q. reckoning are exactly those set forth in the present volume.

[1] Porteus, S. D., *Porteus Tests—The Vineland Revision* (Vineland, N. J., Research Dept. Publications, 1919), No. 16.

[2] Porteus, S. D., *Guide to Porteus Maze Test* (Vineland, N. J., Research Dept. Publications, 1924), No. 25.

[3] Porteus, S. D., *The Maze Test and Mental Differences* (Vineland, N. J., Smith Printing House, 1933).

[4] Porteus, S. D., *The Porteus Maze Test and Intelligence* (Palo Alto, Calif., Pacific Books, 1950).

[5] Porteus, S. D., *The Maze Test: Recent Advances* (Palo Alto, Calif., Pacific Books, 1955).